Modeling, Representing and Learning of Visual Categories

A dissertation submitted to the
TECHNISCHE UNIVERSITÄT DARMSTADT
Fachbereich 20

for the degree of
Dr. ing.

presented by

MARIO FRITZ
Dipl.–Inf.

born 16th of January, 1978
in Adenau, Germany

Prof. Dr. Bernt Schiele, examiner
Prof. Dr. Pietro Perona, co-examiner

Date of Submission: 12th of June, 2008
Date of Defense: 8th of August, 2008

2008
D17

Abstract

This thesis is concerned with the modeling, representing and learning of visual categories for the purpose of automatic recognition and detection of objects in image data. The application area of such methods ranges from image-based retrieval over driver assistance systems for the automotive industry to applications in robotics. Despite the exciting progress that has been achieved in the field of visual object categorization over the last 5 years, we have still a long way to go to measure up to the perceptual capabilities of humans. While humans can recognize far beyond 10000 categories, machines can nowadays recognize only close to 300 categories with moderate accuracy in constraint settings. For more complex tasks the number of categories is a magnitude lower.

Existing approaches reveal a surprising diversity in the way how they model, represent and learn visual categories. To a large extend, this diversity is a result of the different scenarios and categories investigated in the literature. This motivated us to develop methods that combine capabilities of previous methods along these 3 axes: Modeling, Representing and Learning. The resulting approaches turn out to be more adaptive and show better performance in recognition and detection tasks on standard datasets. Therefore, the scientific contribution of this thesis is structured into 3 parts:

Combination of different modeling paradigms One basic difference in modeling is, whether a method models the similarities within one category or the differences with respect to other categories. Since both views have their assets and drawbacks, we have developed a hybrid approach that successfully combines the strength of both approaches.

Combination of different learning paradigms While supervised approaches typically tend to have better performance, the high annotation efforts poses a big obstacle towards a larger number of recognizable categories. Unsupervised methods in combination with the overwhelming amount of data at hand (e.g. internet search) constitute an appealing alternative. Given this background we developed a method which makes use of different levels of supervision and consequently achieves better performance by considering unannotated data.

Combination of different representation paradigms Previous approaches differ strongly in the way they represent visual information. Representations range from local structures over line segments to global silhouettes. We present an approach that learns an effective representation directly from the image data and thereby extracts structures that combine the mentioned representation paradigms in a single approach.

Zusammenfassung

Diese Dissertation beschäftigt sich mit dem Modellieren, Repräsentieren und Erlernen von visuellen Kategorien zum Zweck der automatischen Erkennung und Detektion von Objekten in Bilddaten. Der Anwendungsbereich solcher Methoden erstreckt sich von bildbasierten Suchfunktionen, über Fahrerassistenzsysteme in der Automobilindustrie bis hin zu Anwendungen in der Robotik. Trotz des Fortschritts, den die Forschung gerade in den letzten 5 Jahren in dem Gebiet der visuellen Objektkategorisierung erreicht hat, ist man heute noch weit von den Wahrnehmungsfähigkeiten eines Menschen entfernt. Während Menschen mit Leichtigkeit weit über 10000 Kategorien erkennen, können Maschinen heutzutage nur an die 300 Kategorien mit mäßiger Präzision unter eingeschränkten Bedingungen unterscheiden. Für komplexere Aufgaben ist die Anzahl sogar eine Größenordnung kleiner.

Bestehende Ansätze basieren auf einer erstaunlichen Vielfalt verschiedener Methoden visuelle Kategorien zu modellieren, zu repräsentieren und zu erlernen. Diese Vielfalt ist zum großen Teil ein Resultat der verschiedenen Szenarien und Kategorien die in der Literatur untersucht wurden. Dies motivierte uns Methoden zu entwickeln, die die Fähigkeiten vorangegangener Methoden entlang der 3 Achsen – Modellieren, Repräsentieren und Lernen – kombinieren. Die resultierenden Ansätze zeigen eine höhere Adaptivität sowie verbesserte Performanz in Erkennungs- und Detektionsaufgaben auf standardisierten Datensätzen. Der wissenschaftliche Beitrag dieser Dissertation ist demzufolge in 3 Teile gliedern:

Kombination verschieder Modellierungsparadigmen Ein grundlegender Unterschied in der Modellierung ist, ob eine Methode die Gemeinsamkeiten innerhalb einer Kategorie oder die Unterschiede zu anderen Kategorien modelliert. Beide Sichtweisen haben ihre Vorzüge und Nachteile, weshalb wir einen hybriden Ansatz entwickelten, der die Stärken beider Ansätze erfolgreich kombiniert.

Kombination verschiedener Lernparadigmen Während überwachte Lernverfahren typischerweise bessere Performanz erzielen stellt der Annotierungsaufwand eine große Hürde auf dem Weg zu einer größeren Anzahl von erkennbaren Kategorien dar. Unüberwachte Verfahren in Kombination mit der überwältigenden Menge an verfügbaren Bildern (z.B. Internetsuchmaschinen) sind eine attraktive Alternative. Vor diesem Hintergrund entwickelten wir ein Verfahren, welches verschiedene Stufen der Überwachung des Lernprozesses nutzt und somit unter Hinzunahme der unannotierten Daten eine bessere Performanz erzielt.

Kombination verschiedener Repräsentationsparadigmen Bisherige Ansätze unterscheiden sich stark in der Art und Weise wie visuelle Information repräsentiert wird. Die Repräsentationen reichen von lokalen Strukturen, über Liniensegmente bis hin zu globalen Silhouetten. Wir stellen einen Ansatz vor, der eine effektive Repräsentation direkt von den Bilddaten lernt und dabei Strukturen extrahiert, die die genannten Repräsentationsparadigmen in einem Ansatz kombiniert.

Acknowledgments

I would like to take the opportunity to thank all the people who advised, supported and encouraged me throughout my thesis.

First of all, I would like to thank Prof. Bernt Schiele for being a great advisor. I'm very grateful, for all of his contributions on a professional level as well as his encouragements and patience. I also would like to thank Prof. Pietro Perona for his interest in my work and his valuable comments.

My work at the TU Darmstadt would not have been as pleasant without my wonderful colleagues, whom I would like to thank for numerous collaborations, discussions and their support: Kristof Van Laerhoven, Gyuri Dorko, Nicky Kern, Tam Huynh, Michael Stark, Micha Andriluka, Christian Wojek, Ulf Blanke, Andreas Zinnen, Ulrich Steinhoff, Ursula Paeckel, Maja Stikic, Victoria Carlsson, Stefan Walk. These also include my former master students Nikodem Majer, Paul Schnitzspan and Sandra Ebert with whom I enjoyed working a lot.

Special thanks go to Bastian Leibe and Krystian Mikolajczyk who have been great tutors during the beginning of my phd as well as Edgar Seemann who has been an excellent office mate with whom I had hours of valuable discussions.

I am grateful for the EU Project CoSy that provided funding as well as the opportunity to meet many interesting researchers at the different project sites. Especially, I would like to thank Barbara Caputo and Geert-Jan Kruijff for successful collaborations.

Finally, I would like to thank my parents for their love and support. They gave me a place where I could always return to.

Contents

1 Introduction 1
- 1.1 Contributions . . . 2
- 1.2 Outline . . . 4

2 Related Work on Visual Categorization of Objects 7
- 2.1 Approaches to Visual Categorization . . . 8
  - 2.1.1 Model Paradigm . . . 8
  - 2.1.2 Representation Paradigm . . . 12
  - 2.1.3 Learning Paradigm . . . 14
- 2.2 Methods . . . 15
  - 2.2.1 Implicit Shape Model . . . 15
  - 2.2.2 Support Vector Machines . . . 18
  - 2.2.3 Topic Models . . . 22
- 2.3 Inspiration by Previous Work for the Contributions of this Thesis . . . 24
  - 2.3.1 Generative/Discriminative Hybrid Model for Detection . . . 24
  - 2.3.2 Weakly Supervised Learning by Discovery of Reoccurring Patterns . . . 24
  - 2.3.3 Integrating Different Levels of Supervision in a Cross-Modal Setting . . . 25
  - 2.3.4 Generative Decompositions of Visual Categories . . . 25

3 Integrated Representative/Discriminative Approach 27
- 3.1 Previous Approaches . . . 28
- 3.2 Integrated Approach . . . 29
  - 3.2.1 Generation of an Appearance Codebook . . . 30
  - 3.2.2 Representation in Codebook Coordinates . . . 30
  - 3.2.3 SVM Verification with Local Kernels . . . 31
  - 3.2.4 Discussion . . . 33
- 3.3 Experiments . . . 33
  - 3.3.1 Data . . . 33
  - 3.3.2 Comparison with Original Approaches . . . 34

3.3.3 Multi-Category Discrimination . . . 34
3.3.4 Discriminant Category Detection . . . 37
3.3.5 Results on the PASCAL VOC Challenge 2005 . . . 37
3.4 Conclusion . . . 43

4 Weakly Supervised Learning via Scale-Invariant Patterns 45
4.1 Object Discovery . . . 46
4.1.1 Generic Codebook Representation . . . 46
4.1.2 Scale-Invariant Patterns (SIPs) . . . 47
4.1.3 Method . . . 48
4.1.4 Evaluation . . . 49
4.2 Object Scale Estimation . . . 50
4.2.1 Method . . . 50
4.2.2 Evaluation . . . 53
4.3 Model Estimation for Detection . . . 55
4.4 Experiments . . . 55
4.4.1 Image Ranking . . . 56
4.4.2 Visual Category Detection Task . . . 56
4.5 Conclusion . . . 59

5 Cross-Modal Learning at Different Levels of Supervision 61
5.1 Tutor-driven Learning Scenario . . . 61
5.1.1 Related Work . . . 62
5.2 Vision Sub-System . . . 63
5.2.1 Object Representation by Sparse Scale-Invariant Patterns . . . 64
5.2.2 Object Discovery . . . 65
5.2.3 Unsupervised Visual Grouping . . . 66
5.2.4 A Joint Model for Visual Categorization from Supervised, Semi-Supervised and Unsupervised Input . . . 67
5.3 Dialogue Sub-System . . . 68
5.4 Spatial Reasoning and Cross-Modal Association . . . 70
5.5 Experiments . . . 73
5.5.1 Label Propagation and Conflict Resolution . . . 73
5.5.2 Quantitative Evaluation . . . 74
5.6 Conclusions . . . 75

6 Decomposition of Visual Categories Using Topic Models 77
6.1 Decomposition of Visual Categories . . . 78
6.1.1 Data Representation . . . 78
6.2 Discovery of Visual Categories . . . 79

6.2.1 Unsupervised Google Re-Ranking Task 79
6.2.2 Unsupervised Object Class Discovery 80
6.3 Detection of Visual Categories 81
6.3.1 Generative/Discriminative Training 81
6.3.2 Sliding Window Approach to Detection 84
6.3.3 Speed-ups: Linear Topic Response and Early Rejection 84
6.4 Experiments 85
6.4.1 Efficient Use of Training Examples and Parameter Selection . 85
6.4.2 Comparison of Methods for Estimation and Evaluation of Approximate Inference 86
6.4.3 Comparison to state-of-the-art on PASCAL'06 VOC detection challenge 87
6.4.4 Comparison to shape features on ETH shape database 89
6.5 Conclusions 90

7 Extensions Towards Explicit Multi-View Modeling 91
7.1 Data Cleaning by LDA Likelihood Ranking 91
7.2 Generative Object Decompositions for Multi-Viewpoint Modeling . . 93
7.2.1 Towards the Representation of Multi-View Object Categories . 93
7.2.2 Unsupervised Viewpoint Discovery 97
7.2.3 Bounding Box Regression 98
7.3 Pascal Visual Object Challenge 2007 100

8 Conclusion 103
8.1 Discussion of Contributions 103
8.2 Perspectives 104

List of Figures 111

List of Tables 113

Bibliography 115

Curriculum Vitae 129

Publications 131

1

Introduction

Building a common ground for humans and machines to exchange information is a challenging problem. Making progress in this area would change how we collaborate with machines and how machines can assist us in our tasks. However, automatic interpretation of image data remains a bottleneck despite the exciting progress in recent years. While humans recognize thousands of visual categories with ease, machines lag far behind. Bridging this gap would bring us an important step closer to unleashing the full potential of applications like autonomous robots for domestic scenarios or rescue missions, visual surveillance, driver assistance, content-based image search and vision for the blind - to name just a few.

While there has been good progress in automatic speech and text analysis, image and video data has shown to be notoriously hard to categorize by machines. It is only in last 5 years, that the modeling of visual categories has improved to a level that is becoming increasingly interesting for practical applications (Fergus *et al.*, 2003). More recently, impressive performance has been shown on specific tasks like pedestrian detection (Leibe *et al.*, 2005).

Despite the exciting progress in recent years we are still missing the adaptive representations that are general yet descriptive enough to encode the overwhelming diversity encountered in visual categories. While humans handle more than 10.000 object categories with ease (Biederman, 1987), today's vision systems can recognize less than 300 categories with moderate accuracy in constraint settings (Griffin *et al.*, 2007; Varma and Ray, 2007) and for more challenging tasks like detection the number is a magnitude lower (Everingham *et al.*, 2007).

Figure 1.1 illustrates some of the challenges. As the intra-class variation between the dog examples is high, it is very difficult to generalize from a few examples to the whole class of all dogs. On the other hand, we see that there is quite some similarity to instances of other categories as the presented cow and donkeys. In particular, if we had chosen a texture-based representation that worked well for the task of categorizing materials, it is very likely to fail at classifying these animals correctly.

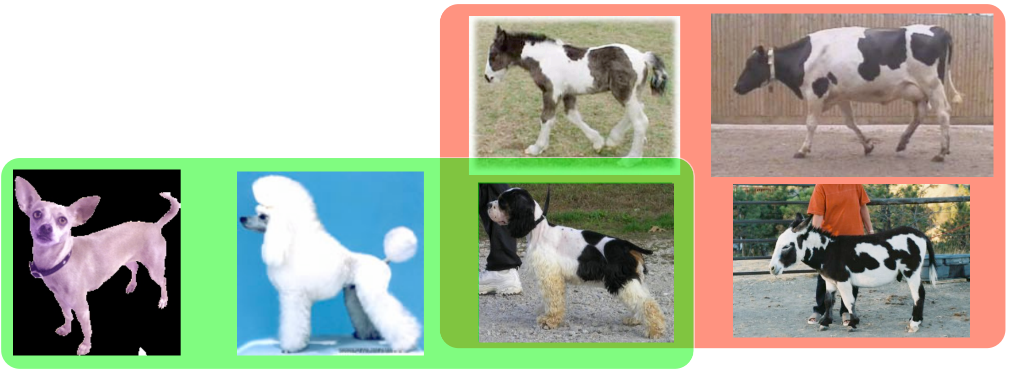


Figure 1.1: Example of high intra-class variation between the dogs and low inter-class variation to instances of other categories.

1.1 Contributions

In Section 2, we sketch 3 main axes along which we categorize different methods proposed in the literature. We are inspired by this topology of related approaches to derive methods that integrate different paradigms along the described axes. We believe that the obtained flexibility and adaptivity is crucial for more robust and scalable systems for visual categorization.

Accordingly, the contributions of this thesis can be grouped into the following 3 groups.

Combining Different Model Paradigms: Hybrid Generative/Discriminative Models for Object Category Detection (Fritz *et al.*, 2005)

- We propose a new approach which tightly integrates a generative with a discriminative approach into a single categorization framework. This tight integration is made possible by a unified data representation used by both approaches. The new integrated approach is beneficial with respect to the initial, probabilistic detector, since the new approach preserves the generalization capabilities but increases its accuracy in rejecting false positives. Since the initial detector effectively acts as a pre-filter to the discriminative part of the algorithm, the integration is also beneficial with respect to the discriminative part by using the discriminative power only where it is needed, namely on visually similar appearances of object classes.

- We present experimental results which show the superiority of the new integrated approach with respect to its building blocks both in terms of detection performance and of a significant reduction of false positives on challenging databases. The new approach also outperforms state-of-the-art object categorization methods on challenging multi-scale data sets.

- We show that the integrated approach improves over and extends the original discriminative model in various respects: the new approach is scale invariant, enables localization of the object in the scene, and allows cross-instance learning of object category models.

Combining Different Learning Paradigms:

- *Weakly Supervised Learning for Accurate Category Detection (Fritz and Schiele, 2006)*
  - We propose a novel scheme to discover object category instances in images. Our approach is based on the idea to estimate the locations and scales of reoccurring patterns. The estimates can be seen as an automatic annotation procedure of the training data.
  - We experimentally show the applicability of this idea of reoccurring structure for object discovery for several object classes.
  - We show how to use the estimated annotations to learn object class models for object detection and localization.
  - We analyze the performance of such object class models on standard datasets. Most interestingly, we even surpass our supervised baseline by adding more unlabeled training examples.
- *Combining Different Levels of Supervision in a Cross-Model Learning Scenario (Fritz et al., 2007)*
  - We proposes a novel method that uses unsupervised training to obtain visual groupings of objects and a cross-modal learning scheme to overcome inherent limitations of purely unsupervised training.
  - The method uses a unified and scale-invariant object representation that allows to handle labeled as well as unlabeled information in a coherent way. One of the potential settings is to learn object category models from many unlabeled observations and a few dialogue interactions that can be ambiguous or even erroneous.
  - First experiments demonstrate the ability of the system to learn meaningful generalizations across objects already from a few dialogue interactions.

Combining Different Feature Paradigms: Adaptive representations based on generative decompositions (Fritz and Schiele, 2008) The main focus of this part is a new object representation that aims to combine different feature paradigms to make a step towards more scalable object representations applicable to a wide range of objects and suited both for unsupervised as well as supervised learning.

- We present a novel approach that allows to learn a low-dimensional representation of object classes by building a generative decomposition of objects. These learned decompositions of objects range from local appearance features to global silhouette-like features shared across object classes. This generative model of objects is directly applicable to unsupervised learning tasks such as visual object class discovery.

- We combine the low-dimensional and generative decomposition of objects with a discriminative learning framework to enable supervised training and competitive object class detection.

- We present a series of experiments which show the properties of the approach (local vs. global features, feature sharing, unsupervised vs. supervised learning) and compares the approach with the state-of-the-art. Interestingly, the approach outperforms both unsupervised techniques as well as supervised techniques on various tasks on common databases.

1.2 Outline

This thesis is structured as follows:

Chapter 2: Related Work on Categorization of Visual Categories We put this thesis in the context of the related work by organizing related approaches by the different modeling, representation and learning paradigms they use. After reviewing 3 methods which are used as building block in this thesis, we detail how the different parts of this thesis are inspired by previous work.

Chapter 3: Integrated Representative/Disciminative Approach This chapter describes our efforts to combine different modeling paradigms. In particular, a generative model is combined with a discriminative classifier to build a hybrid model.

The work represented in this chapter corresponds to the ICCV'05 publication (Fritz *et al.*, 2005): "Integrating Representative and Discriminant Models for Object Category Detection"

Chapter 4: Weakly Supervised Learning via Scale-Invariant Patterns We propose a method to discover object category instances in image databases as reoccurring patterns in a weakly supervised fashion. The results show that we can bridge the gap between image level annotation and highly supervised detection performance obtained by pixel level annotation.

The work represented in this chapter corresponds to the DAGM'06 publication (Fritz and Schiele, 2006): "Towards Unsupervised Discovery of Visual Categories"

Chapter 5: Cross-Model Learning at Different Levels of Supervision We describe a system that represent our efforts towards integrating different levels of supervision in one consistent framework. The capabilities are demonstrated in a cross-modal, interactive learning scenario that related utterances of a tutor with the visual features and relations in a scene.

The work represented in this chapter corresponds to the ICVS'07 publication (Fritz *et al.*, 2007): "Cross-Modal Learning of Visual Categories using Different Levels of Supervision"

Chapter 6: Decomposition of Visual Categories Using Topic Models We present a method for learning representations as generative decompositions of object category instances. This approach manages to combine different representation paradigms ranging from local to global features.

The work represented in this chapter corresponds to the CVPR'08 publication (Fritz and Schiele, 2008): "Decomposition, Discovery and Detection of Visual Categories Using Topic Models"

Chapter 7: Extensions Towards Explicit Multi-View Modeling Finally, we extend the results from the previous chapter to handle multi-view data in a more explicit manner. In particular, we provide model introspection how different view-points are represented and how we can deal with changing aspect ratios.

2

Related Work on Visual Categorization of Objects

In the context of computer vision, research in visual categorization has focused mostly on basic-level categories. These are categories that - although embedded in a hierarchy of more abstract and specific categories - are more likely to be used by a human (Lakoff, 1987; Rosch *et al.*, 1976). There is some bias humans have towards using these categories. For example, when presented with an image of a cat, the observer is more likely to name it as a cat than an animal, quadruped or siamese cat. Throughout this thesis we will be concerned with such basis-level categories, although there are no inherent limitation to apply the presented methods to different levels of categorization.

For the past decades, research progress was most notably in the domain of recognizing specific objects Murase and Nayar (1995); Schiele and Crowley (2000). There were only a view exceptions that extended to more variable object classes like digits (e.g. LeCun *et al.* (1998)), faces (e.g. Turk and Pentland (1991); Murase and Nayar (1995); Cootes *et al.* (1998); Viola and Jones (2001), cars (e.g. Schneiderman and Kanade (2000)) and humans (e.g. Gavrila and Davis (1996)).

Only recently there has been significant progress in more general approaches to visual category modeling (e.g. Burl and Perona (1996)), learning (e.g. Fergus *et al.* (2003)), robustness (e.g. Leibe *et al.* (2005)) and scalability to more classes (e.g. Varma and Ray (2007)).

In order to set this thesis into the context of previous work Section 2.1 gives an overview of relevant approaches. We structure the related work along 3 axes which represent the different modeling, representation and learning paradigms encountered in the literature. After reviewing 3 methods in Section 2.2 that are used in this thesis, we describe in Section 2.3 how the previous approaches inspired the work of this thesis – in particular with respect to the 3 axes modeling, representing and learning.

2.1 Approaches to Visual Categorization

Despite the exciting progress, there is a surprising diversity in the proposed approaches. This could be interpreted as a lack of maturity of the research area. More likely it is an artifact of the diverse experimental scenarios investigated in the literature. Features which work well on cars don't necessarily perform so well on leopards.

In the following, this section sketches 3 main axes along which we categorize different methods proposed in the literature. We are inspired by this topology of related approaches to derive methods that integrate different paradigms along the described axes. We believe that it is important to discuss and integrate these paradigms to obtain adaptive and flexible approaches that lead to more robust and scalable systems for visual categorization.

2.1.1 Model Paradigm

There is a fundamental difference in models whether they describe what all category instances have in common (*generative model*) or what distinguishes them form the other categories (*discriminative model*). In technical terms this boils down to the design decision if the data X associated with a training example is explicitly modeled or if the approaches focus on the class label y. Often generative models are also associated with probabilistic models, as they allow for sampling from the model which exploits the generative nature. However, this is not correct, as the class posterior trained in a discriminative fashion can also be modeled by a probability density function $p(y|X)$. These models are often termed *conditional models*. Following common practice, we will refer to them also as *discriminative models*. More frequently people retreat to a statistical machine learning framework and seek just a discriminant function $y = f(X)$ which performs best on the prediction task without worrying about probabilistic modeling. These kind of models are commonly called *discriminant models*. For a more detailed discussion on the spectrum between generative and discriminative and the associated terminology, we refer to Jebara (2002).

Due to the many implication of choosing one of these two statistical paradigms, various approach for visual category recognition and detection have been proposed, ranging form generative, probabilistic models like Burl and Perona (1996); Fergus *et al.* (2003); Leibe *et al.* (2004) to discriminant ones likeTorralba *et al.* (2004); Viola and Jones (2004); Nilsback and Caputo (2004); Dalal and Triggs (2005). Boosting (Freund and Schapire, 1995) and maximum margin classification in a Support Vector Machine framework (Schölkopf and Smola, 2001) are the most prominent methods for discriminant modeling.

In the following we will provide an overview of the different arguments brought forward from the different schools of thought.

Benefits of generative models

- *More principled way of dealing with missing data (e.g. occlusion) (Holub* et al.*, 2008)*: In particular in the case of probabilistic models, missing values can be dealt with by marginalizing over the missing variable. This leads still to sensible prediction within the formulated model. For a discriminant function, there is no canonical way to deal with such artifacts.

- *Better performance for small training sets (Holub* et al.*, 2008; Ng and Jordan, 2002; Fei-Fei* et al.*, 2003a)*: The argument is mostly based on the fact that probabilistic models can easily incorporate prior information via a bayesian framework, that supports learning from small sample sizes by modeling prior expectations on the parameters like in Fei-Fei *et al.* (2003a) or transferring knowledge from other models like in Bart and Ullman (2005). However Ng and Jordan (2002) present a more detailed analysis on non-vision databases that shows a more general improvement of the generative approach due to the model bias (which then becomes a disadvantage in the limit when having lots of training examples).

- *Modeling expert knowledge (Holub* et al.*, 2008)*: Graphical models provide an excellent tool to describe and visualize the conditional independence structure of the model (Bishop (2007)). This is one way to incorporate expert knowledge about the latent structure into a model. Furthermore, priors provide a convenient and consistent way to model expectations on parameters and unobserved variables.

- *Learning one category at a time improves scalability (Holub* et al.*, 2008)*: Generative approaches model what's common between the presented instances. It is ignorant about the differences to other categories in contrast to the discriminant approaches. By not exploiting the discriminant information, the generative models can be learnt independently. This inherent parallelism can lead to better scalability. Learning the pair-wise differences between classes is a powerful approach, but can lead to high computational costs (Varma and Ray, 2007).

- *Dealing with deformation and intra-class variation*: Generative models can tolerate significant intra-class variation of object appearance and deformations like in Fergus *et al.* (2003); Leibe *et al.* (2004); Felzenszwalb and Huttenlocher (2005). However, the price for this robustness typically is that they tend to produce a significant number of false positives. This is particularly true for object classes which share a high visual similarity such as horses and cows. Discriminative methods typically cope with these issues by increased training set sizes of virtual (Decoste and Schölkopf, 2002) or real examples (Dalal and Triggs, 2005).

- *Incremental Learning*: Generative properties have been used for incremental learning in a Bayesian settings in Fei-Fei *et al.* (2004) as well as for subspace methods of Skočaj and Leonardis (2003). However, also discriminative approaches were combined with heuristics to facilitate incremental learning for SVMs in Cauwenberghs and Poggio (2001); Luo *et al.* (2007) as well as boosting in Opelt *et al.* (2006).

- *Hierarchical structures*: There is a long tradition in hierarchical modeling in the statistics literature (e.g. Gelman *et al.* (2004)). In particular, the modular construction of composed solutions to complex problems lends itself to hierarchical design as demonstrated in Sudderth *et al.* (2005); Bouchard and Triggs (2005). Hierarchical models are considered to be one key ingredient towards scalable approaches (Fidler and Leonardis (2007)).

- *More explicit model assumptions (Friedman, 1997)*: Generative model tend to make the model assumptions more explicit. In particular parametric models presented in graphical structures translate directly to a generative process Sudderth *et al.* (2005). Hence distribution and independency assumptions can be verified separately.

Benefits of discriminative models

- *Higher precision*: Ng and Jordan (2002) have illustrated that discriminative models tend to have a smaller asymptotic error. In particular, discriminative models are optimized for the classification task and don't care about the data distribution. This typically results in increased performance (Friedman, 1997; Ulusoy and Bishop, 2005).

- *Easier task*: Vapnik (1996) argues that estimating the distribution of X is often an unnecessary overhead that decreases performance. A direct, discriminative mapping form the data X to the labels y can therefore make more efficient use of the data.

- *Simpler and faster to evaluate*: Discriminative classifiers tend to be faster to evaluate (Ulusoy and Bishop, 2005). Many classifiers have a deterministic, feed-forward structure in contrast to often costly inference techniques required for generative, probabilistic models (Bishop (2007)).

- *Feature selection* This allows for example to explicitly learn the discriminant features of one particular class vs. background (Viola *et al.*, 2003; Dorko and Schmid, 2003) or between multiple classes (Torralba *et al.*, 2004; Nilsback and Caputo, 2004).

- *Flexible decision boundaries*: Discriminative methods enable the construction of flexible decision boundaries, resulting in classification performances often superior to those obtained by purely probabilistic or generative models (Jaakkola and Haussler, 1999; Ng and Jordan, 2002). There are two prominent reasons for this effect mentioned in the literature. First, generative models are often derived from probabilistic approach that have some bias due to distribution assumption and prior models. This leads to a higher error given a certain amount of data when compared to discriminative approaches as shown in Ng and Jordan (2002). However, this argument wouldn't be valid for a method based on non-parametric density estimated in a maximum likelihood fashion. Second, focusing on the estimation of a decision boundary doesn't "waste" samples on estimating data properties and areas not relevant to the classification task - in particular areas far away from the decision boundary. As a conclusion, discriminative models are more likely to obtain a better estimate of a complex decision boundary by neglecting task-irrelevant properties of the data. For a more detailed discussion and quantitative results, we refer to Friedman (1997).

Hybrid methods combining generative and discriminant approaches While so far the object recognition community has in most cases chosen one of these two modeling approaches, there has been an increasing interest in the machine learning community in developing algorithms which combine the advantages of discriminative methods with those of probabilistic generative models (e.g. Jaakkola and Haussler (1999)).

In the following, we outline the different approaches that were investigated in the vision community to combine the merits of both paradigms. We propose the following grouping of the diverse methods:

- *Discriminative training/optimization of generative model*: In Hillel *et al.* (2005) a generative model with a star-like topology is trained. To improve the performance the influence of the parts undergoes an additional discriminative optimization in order to improve discrimination between categories. A similar idea is explored in Li *et al.* (2005) where an initial generative stage identifies relevant feature component on which the final classifier is trained.

- *SVM learning using kernels on probabilistic models*: The following approaches estimate a probability density function (pdf) on the input data and use the obtained statistic in a specialized kernel function to train a discriminant model. Possible choices for such kernels are the Fisher kernel (Jaakkola and Haussler (1999)) and the Kullback Leibler kernel (Moreno *et al.* (2005)). For a more general overview on the topic of probability product kernels we refer to Jebara *et al.* (2004).

    - *Fisher kernel*: The fisher kernel (Jaakkola and Haussler (1999)) measures similarity between two examples by the scalar product between the likelihood gradients with respect to the model parameters. Therefore the underlying probabilistic model has to be trained beforehand in order to estimate the maximum likelihood parameters.

  In Holub *et al.* (2005) this kernel is employed for visual recognition by building on the constellation model (Fergus *et al.*, 2003) as the generative part.
  - *Kullback Leibler Kernel*: In Vasconcelos *et al.* (2004) the Kullback Leibler kernel is used to discriminate between bag-of-word representations. The main motivation in their work is not the generative/discriminative aspect, but to define a kernel between sets of local features. To achieve this, the sets of local features are modeled as samples from gaussian distributions. The kernel is then defined as Kullback Leibler divergence between these distributions.

- *Discriminant hypotheses verification*: Based on initial detection by the generative detection system presented in Leibe *et al.* (2004) an additional discriminant verification stage based on chamfer matching is proposed in Leibe *et al.* (2005). In particular as the initial detections are based on a star-model with conditional independent part evidences, the global chamfer matching approach adds additional information to the detection process.

- *Discriminative feature selection and weighting*: Most approaches mentioned so far applied a discriminative approach on top of the generative one. Methods which do features selection (Dorko and Schmid, 2003) or feature weighting (Mikolajczyk *et al.*, 2006) go the opposite way. The discriminance of single features is assessed in order to weight or even fully reject non-informative features.

- *Scene reasoning*: In Tu *et al.* (2003) a generative model for segmenting scenes is combined with a object detector trained in a discriminant fashion. The combination turns out to be beneficial for image parsing into regions and objects. The discriminant detections are treated as additional evidence for the generative model.

- *Generative synthesis and rendering*: Finally, we want to pay credit to approaches that employ generative aspects in a non-probabilistic way. They employ rendering techniques to generate virtual examples that improve generalization of the discriminative approach by increasing the training set (Decoste and Schölkopf, 2002; Everingham and Zisserman, 2005; Chiu *et al.*, 2007). In terms of efficiency, approach like Decoste and Schölkopf (2002) or Kapoor *et al.* (2007) are very interesting as they have developed ways to predict which virtual examples are most informative and therefore should be considered next.

2.1.2 Representation Paradigm

Feature representations based on gradient histograms have been popular and highly successful. The proposed features range from local statistics like SIFT (Lowe, 2004) to global representations of entire objects (Gavrila, 1998) and from sparse interest points Mikolajczyk and Schmid (2005) to dense features responses (Dalal and Triggs, 2005).

In the following we will outline the different choices and the associated benefits and drawbacks:

Local representation Local feature representations are very popular and lead to a series of successful approaches for visual categorization like Burl and Perona (1996), Leibe *et al.* (2008), Fergus *et al.* (2003) and Mikolajczyk *et al.* (2006). In particular the combination with probabilistic models lead to approaches with high recall due to robustness with respect to occlusion and the ability to capture the variance of visual categories well. The drawback of breaking down the images into jigsaw puzzels is that the models which describe the possible constellations are often weakly structured (e.g. bag-of-words as in Csurka *et al.* (2004)), simplifying (e.g. ISM model with star topology in Leibe *et al.* (2008); the missing interdependencies between parts can lead to hallucination of superfluous parts that correspond to fake evidence) or expensive to evaluate (e.g. constellation model in Fergus *et al.* (2003)).

Global representation A prominent example for global presentations is chamfer matching of silhouettes (Gavrila, 1998). In contrast to the local representation, global consistency is always verified. On the other hand, these approaches typically loose recall due to the inflexible structure - unless provided with large amounts of training data.

Semi-local representation Mohan *et al.* (2001) proposed to detect part structures by more local histograms first which are combined to object detections afterwards. In order to not define the sub-parts manually, Laptev (2006) constructs a representation on random sub-crops. This is an extension of the method of Levi and Weiss (2004). A boosting method is used to select a representation most suited for the task.

Also shape based approaches have recently received a lot of attention, as they provide a first abstraction level form the image gradients and they are capable to represent elongated structures as edges as well as more localized features like corners. In particular, progress in the edge detection procedure (Martin *et al.*, 2004) and description of edge structures (Ferrari *et al.*, 2008) revived the discussion.

Sparse features Sparse feature representation are based on methods to select distinctive points in the image. Only these selected points are represented and considered for further processing. Typically, the selection is based on local maxima detection of an interest function that responds to edge or corner-like structures (Mikolajczyk and Schmid, 2005). Besides achieving data reduction, the development of scale-invariant interest points (Lindeberg, 1998) spurred the success of these approaches.

Dense representation Studies have shown that sparse representations are inferior to a denser sampling in terms of system performance (Nowak *et al.*, 2006). While the difference might by less pronounced for some categories, objects where the interest point detector fails to capture a sufficient statistic are better captured by random sampling or sampling on a regular grid.

In particular as machine learning methods made good progress in handling large amounts of data, dealing with noise and extracting the relevant information, combination with dense feature representation have led to state-of-the-art recognition and detection approaches like Dalal and Triggs (2005).

2.1.3 Learning Paradigm

Over the years various approaches have been proposed for the recognition of object categories often based on models learned directly from image data. The approaches, however, vary greatly in the amount of supervision provided for the training data. The types of annotation varies from pixel-level segmentations (e.g. Leibe *et al.* (2008)), over bounding-box annotations (e.g. Viola and Jones (2001)) and image level annotation (e.g. Fergus *et al.* (2003); Winn and Jojic (2005)) to unsupervised methods (e.g. Weber *et al.* (2000); Sivic *et al.* (2005); Fergus *et al.* (2005a)) which do not even require the information which category is presented in which image. While approaches using more supervision tend to require less training data, there is a clear desire to use less supervision typically at the price to use more unlabeled training data.

Supervised learning Traditionally, providing more annotation information results in better performance given the same amount of training data. Besides bounding box information (Viola and Jones, 2001), approaches have shown to successfully exploit pixel-wise segmentations (Leibe *et al.*, 2008) or view-point annotations (Chum and Zisserman, 2007) to increase performance.

Weakly supervised learning Weakly supervised learning typically denotes learning from an image-level annotation of the presence or absence of the object category. Learning object models in this fashion may be formulated in one EM-loop as in e.g. Weber *et al.* (2000). In this method, appearance and structure are learned simultaneously making the learning computationally expensive and thus restricting the complexity of the model. More recently, weakly supervised approach were derived by finding frequent item sets (Quack *et al.*, 2007) or substructure mining (Nowozin *et al.*, 2007) leading to more computational efficiency.

Unsupervised learning Unsupervised learning is very attractive, as the annotation effort and biases introduced by the human labeling process become increasingly problematic for large training sets with many classes (Ponce *et al.*, 2006). As today's internet based services provide image data in abundance, learning in an unsupervised fashion provides a promising alternative to these problems. As those data sources typically return lots of unrelated images (Fergus *et al.*, 2005a) and images of poor quality, these method have to be robust against outliers and a large variety of image degradations (e.g. Fergus *et al.* (2005a); Li *et al.* (2007); Schroff *et al.* (2007)). Recently a variety of approaches have been proposed that are based on topic models such as pLSA (Hofmann, 2001) or LDA (Blei *et al.*, 2003b). Since the underlying model is a bag-of-word representation, the object discovery is based on local appearance alone neglecting structural information (Sivic *et al.*, 2005). E.g. Fergus *et al.* (2005a); Russell *et al.* (2006); Cao and Fei-Fei (2007) extends the initial approach to also include some structural information on top of the pLSA model.

Semi-supervised learning Semi-supervised learning has recently gained a lot of interest in the machine learning literature, as it combines unsupervised approach with supervised information to overcome the inherent limitation of fully data driven approaches. For an overview we refer to Zhu (2005) and Chapelle *et al.* (2006). In the vision community, the impact of these approaches has been less prominent up to now. But approaches like Holub *et al.* (2008) and in particular extension towards active learning like in Kapoor *et al.* (2007) seem promising.

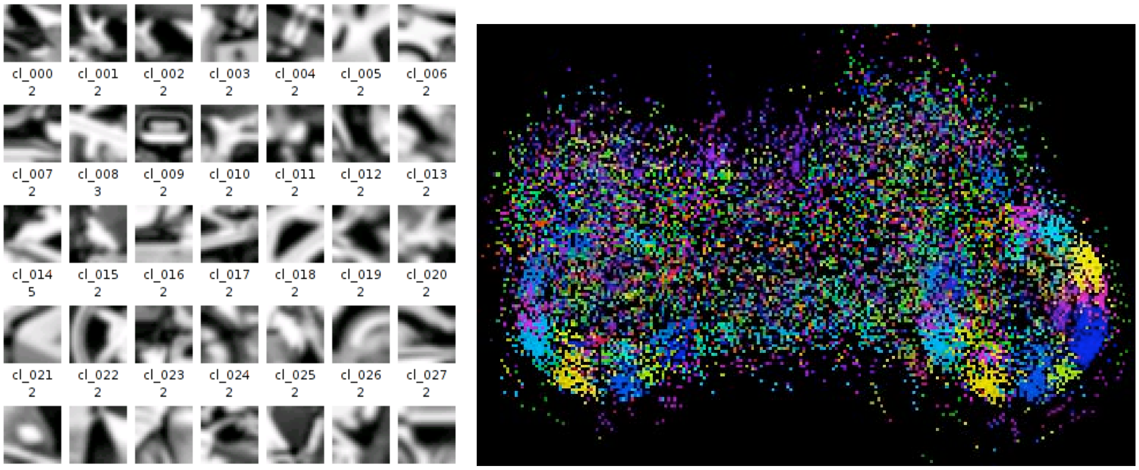
2.2 Methods

This section will outline models from previous work, that this thesis builds on. In particular, this section describes the Implicit Shape Mode (ISM) of Leibe *et al.* (2008) for generative object detection, generative topic models from the text processing domain as in Hofmann (2001) and Blei *et al.* (2003b) as well as discriminant classification by *Support Vector Machines (SVMs)* as described in Schölkopf and Smola (2001).

2.2.1 Implicit Shape Model

The *Implicit Shape Model (ISM)* (Leibe *et al.*, 2008) is a versatile generative, framework for scale-invariant detection of object categories, which has shown good performance on challenging detections tasks. It uses a flexible non-parametric representation for modeling visual object categories by spatial feature occurrence distributions with respect to a visual codebook. In this section we describe how the visual codebook is computed, how the spatial layout of the objects is captured and how the stored information is finally used for detection tasks.

Generation of an Appearance Codebook In the first step, a category-specific appearance codebook is generated. A scale-invariant interest point operator like e.g. Difference-of-Gaussians (DoG) (Lowe, 2004) is applied to all training images



(a) Sample cluster centers of visual codebook.

(b) Example occurrence distribution for motorbikes. The colors correspond to the different clusters the features get matched to.

Figure 2.1: The two key ingredients of the Implicit Shape Model: (a) visual codebook vocabulary and (b) spatial occurrence distribution of the codebook entries. Both statistics are computed on the Caltech motorbike dataset.

and extract image patches with a radius of 3σ of the detected scale. All extracted patches are then rescaled to a uniform size (in our case 25×25 pixels) and grouped using an agglomerative clustering scheme. The patches can be represented by simple normalized gray-values or more sophisticated representations like SIFT (Lowe, 2004) or shape context (Belongie *et al.*, 2002) that typically improve the results as shown in Seemann and Schiele (2005). The resulting clusters form a compact representation of local object structure. In the following, we keep only the cluster centers $C = (\vec{c}_1, \ldots, \vec{c}_R)$ as codebook entries.

Example entries of a codebook trained on motorbike images are shown in Figure 2.1(a). The shown patches are representative for each cluster center, that are obtained by averaging over all cluster members. While some entries might resemble motorbike parts, this is a purely data driven approach that groups patches by visual similarity only.

Non-parametric spatial occurrence distributions For each codebook entry, we then learn its spatial occurrence distribution on the object category with respect to a common center for all motorbikes. The center is either defined by the center of the bounding box annotation or the center of mass when a segmentation mask is provided. Therefore, we perform a second iteration over all training images, again extracting patches around interest points, and record for each $\vec{c}_i$ all locations where it can be matched to the extracted patches. This results in a non-parametric cluster entry occurrence distribution that captures the spatial layout of the category of interest.

An example of such an occurrence distribution for motorbikes is displayed in Figure 2.1(b). The cluster to which the feature matched is color-coded. In particular on the wheels colored areas indicate that specific codebook entries get matched to certain segments of the wheels. There is additional information on the detected feature scale for each feature which is note visualized in this figure.

Detection In order to generate detection hypotheses with locations and scales on test images, we use a scale-invariant version of the ISM approach from Leibe *et al.* (2004). The approach starts by applying the same feature extraction procedure as before. Each patch is matched to the codebook, and matching codebook entries cast votes for possible object positions and scales according to their learned spatial probability distribution. It is important to note that each matched feature cast votes independently of the other features. Although this independence assumption might seem rather crude, it leads to very efficient learning and evaluation of the model as well as to high generalization capability across training instances.

The voting procedure is formalized as follows. Let $\vec{e}$ be an image patch observed at location ℓ. Each matching codebook entry $\vec{c}_i$ generates probabilistic votes for different object categories o_n and locations $\lambda = (\lambda_x, \lambda_y, \lambda_s)$ according to the following marginalization:

$$P(o_n, \lambda | \vec{e}, \ell) \;=\; \sum_i P(o_n, \lambda | \vec{c}_i, \ell) p(\vec{c}_i | \vec{e}) \tag{2.1}$$

where $p(\vec{c}_i|\vec{e})$ denotes the probability that $\vec{e}$ matches to $\vec{c}_i$, and $P(o_n, \lambda|\vec{c}_i, \ell)$ describes the stored spatial probability distribution for the object center relative to an occurrence of that codebook entry. For describing the matching probability, we make the assumption that an image patch can be approximated by the mean of the closest-matching codebook entries $C^*_{\vec{e}} = \{\vec{c}^*_i | sim(\vec{c}^*_i, \vec{e}) \geq \theta\}$, thus setting $p(\vec{c}^*_i|\vec{e}) = \frac{1}{|C^*_{\vec{e}}|}$. Object hypotheses are found as maxima in the 3D voting space using Mean-Shift Mode Estimation (Comaniciu and Meer, 1999) with a scale-adaptive *balloon density estimator* (Comaniciu *et al.*, 2001) and a uniform ellipsoidal kernel K:

$$\hat{p}(o_n, \lambda) = \frac{1}{nh(\lambda)^d} \sum_k \sum_j p(o_n, \lambda_j | \vec{e}_k, \ell_k) K\left(\frac{\lambda - \lambda_j}{h(\lambda)}\right),$$

where n is the total number of features. Once a hypothesis has been found, the contributing votes are backprojected to determine which local features and codebook activations supported it. The original ISM approach additionally computes a full top-down segmentation of the object, which has been shown to improve the results considerably. This is done by back-projecting the support of the hypotheses to infer figure-ground segmentation masks and performing an MDL-based reasoning to resolve multiple and ambiguous hypotheses (Leibe *et al.*, 2004). However, the generation of an object specific visual codebook and the MDL-based reasoning step require figure-ground segmentations for the training images which introduce high annotation effort. For further details we refer the reader to Leibe *et al.* (2008).

2.2.2 Support Vector Machines

Support vector machines (SVMs) (Vapnik (1996), Schölkopf and Smola (2001)) are discriminative models and have recently raised a lot of interest because of their well-founded theoretical background and excellent classification performance across many tasks. In order to describe SVMs, the concept of linear discrimination is reviewed first, which is extended to optimal separating hyperplanes and soft margin hyperplanes. Finally, we will arrive at robust non-linear classification by introducing kernels.

Linear Discrimination The basic idea of a linear decision function is to specify a hyperplane in the input space which separates two classes. Such a hyperplane is defined by the normal form:

$$\vec{w}^T\vec{x} + b = 0 \ , \tag{2.2}$$

where $\vec{w}$ is the direction normal to the hyperplane and $|b|$ is the distance of the hyperplane to the origin of the coordinate system.

For each data sample $\vec{x}$ the distance to this hyperplane can be computed by projection on the vector $\vec{w}$. The sign of the projection tells us on which side of the plane the sample lies. Therefore the decision function is given by:

$$f(\vec{x}) = \mathrm{sign}(\vec{w}^T\vec{x} + b) \tag{2.3}$$

Optimal Separating Hyperplane The solution to the formulation in Equation (2.3) is often not unique. There can be infinitely many hyperplanes that successfully classify the data in the case of linearly separable data. To get to a unique representation, one defines the optimal separating hyperplane to be the one which maximizes the margin to the data samples. In order to compute this plane the following optimization problem is used:

$$\max_{\vec{w},b,||\vec{w}||=1} \{ \min_{i=1,\dots,l} (y_i(\vec{x}_i^T\vec{w} + b)) \} \tag{2.4}$$

This situation is illustrated in Figure 2.2. By normalizing with the length of w, this can be reformulated without the constraint $||\vec{w}|| = 1$.

$$\min_{\vec{w},b} \tfrac{1}{2}||\vec{w}||^2 \tag{2.5}$$

$$\text{subject to} \quad y_i(\vec{x}_i^T\vec{w} + b) \geq 1 \ , \ i = 1, \dots, l \tag{2.6}$$

This is a constrained, convex optimization problem. As a conclusion only one global minimum exists and the solution can be found efficiently. We rewrite the problem by taking the constraints into account via Lagrangian multipliers, obtaining the Lagrange function:

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{l} \alpha_i(y_i(\vec{x}_i^T\vec{w} + b) - 1) \ i = 1, \dots, l \ , \tag{2.7}$$

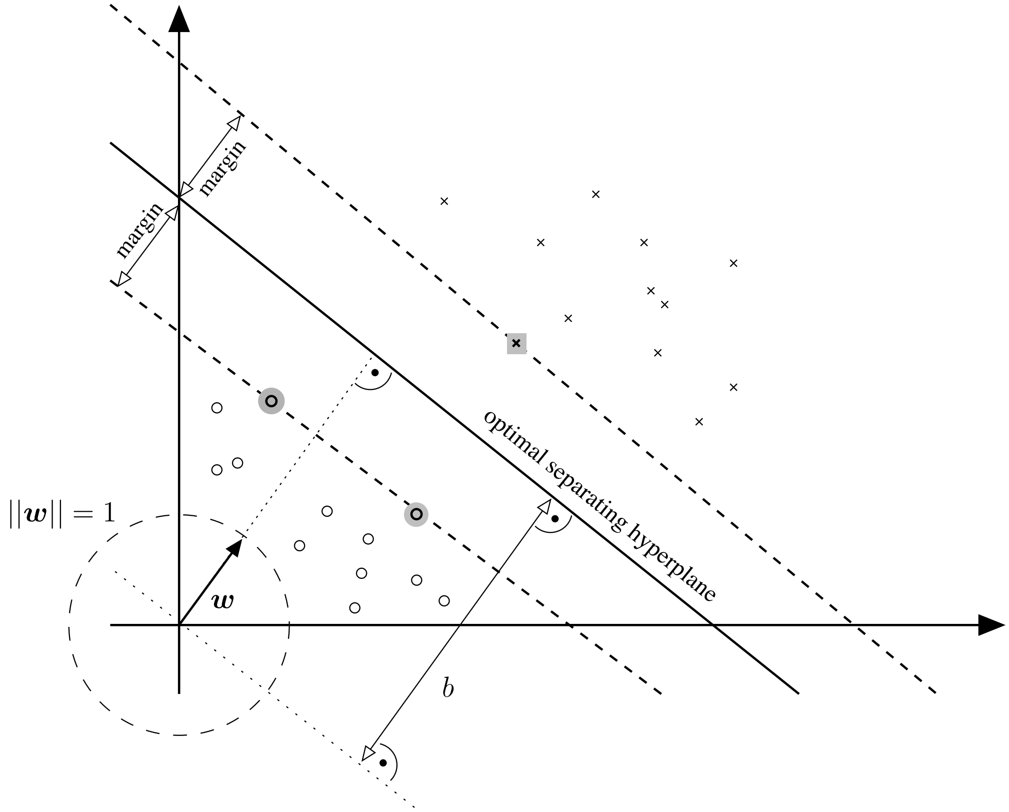


Figure 2.2: Illustration of the normal form of a hyperplane. The orientation of the hyperplane is given by the orthogonal vector $\vec{w}$ which is constraint to $||\vec{w}|| = 1$ and the distance from the origin by b. The hyperplane is chosen to separate the two shown classes with a maximal margin. The so-called support vectors are marked in grey.

where $\vec{\alpha} = \alpha_1, \ldots, \alpha_l$ are the Lagrange multipliers. Therefore the minimum of the optimization problem must satisfy:

$$\frac{\partial}{\partial b} L(\vec{w}, b, \alpha) = 0 \tag{2.8}$$

and

$$\frac{\partial}{\partial \vec{w}} L(\vec{w}, b, \alpha) = 0 \tag{2.9}$$

Substituting eqn. (2.8) and eqn. (2.9) in eqn. (2.7) one obtains the dual optimization problem which is formulated in the variables $\alpha_1, \ldots, \alpha_l$:

$$\max_{\vec{\alpha} \in \mathbb{R}^l} W(\vec{\alpha}) = \textstyle\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{j=1}^{l} \sum_{i=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \tag{2.10}$$

$$\text{subject to} \quad \alpha_i \geq 0 \; \forall i = 1, \ldots, l \text{ and } \textstyle\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.11}$$

As a solution we obtain values for the α which leads to the following decision function:

$$f(\vec{x}) = \operatorname{sign}(\sum_{i=1}^{l} y_i \alpha_i \vec{x}^T \vec{x}_i + b) \tag{2.12}$$

Computing the α on typical data sets it turns out that many of α are 0 and therefore do not contribute to the decision function. The $\vec{x}_i$ with non-zero α are called support vectors. In order to be able to evaluate the model, they have to be stored together with the α and determine the memory size of the model. In Figure 2.2 the support vectors are marked in grey.

The technique reviewed above solves the problem where the data are linearly separable. Frequently, data we encounter in real-world applications do not show this property. But even if that can be solved, additional problems arise from noise, that can lead to wrong and too complex boundaries. Therefore two extensions were introduced (Vapnik, 1996; Schölkopf and Smola, 2001). First, the *kernel trick* extends the linear separable case to more complex problems by employing a non-linear transformation of the data and the *soft margin hyperplane* which can handle noisy data by introducing slack variables with a penalty term.

Soft Margin Hyperplane To account for data samples that cause the data set to be non-separable, *slack variables* $\xi_i \geq 0$ were introduced changing the constraints to:

$$y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i \tag{2.13}$$

To penalize mis-classification as allowed by the slack-variables, an additional penalty term is added which weights the sum over all slack variables ξ_i with the parameter C:

$$\min_{\vec{w},b} \tfrac{1}{2}||\vec{w}||^2 + C \sum_{i=1}^{l} \xi_i \tag{2.14}$$

$$\text{subject to} \quad \xi_i \geq 0\ ,\ y_i(\vec{x}_i^T \vec{w} + b) \geq 1 - \xi_i\ ,\ i = 1, \ldots, l \tag{2.15}$$

In Figure 2.3 the slack variables are illustrated. The solution to this optimization problem is obtained analogously to the linearly separable case (Vapnik, 1996; Schölkopf and Smola, 2001). There is no canonical way to chose the parameter C. It has to be chosen appropriately depending on the task.

Kernel Trick The kernel trick is a method to make a linear classifier more flexible and therefore applicable to more complex problems. The data is transformed into a *feature space* $\mathcal{H}$ by a non-linear mapping Φ which increases the separability of the data:

$$\Phi : \quad \mathbb{R}^d \longrightarrow \mathcal{H} \tag{2.16}$$

$$\vec{x} \longmapsto \Phi(\vec{x}) \tag{2.17}$$

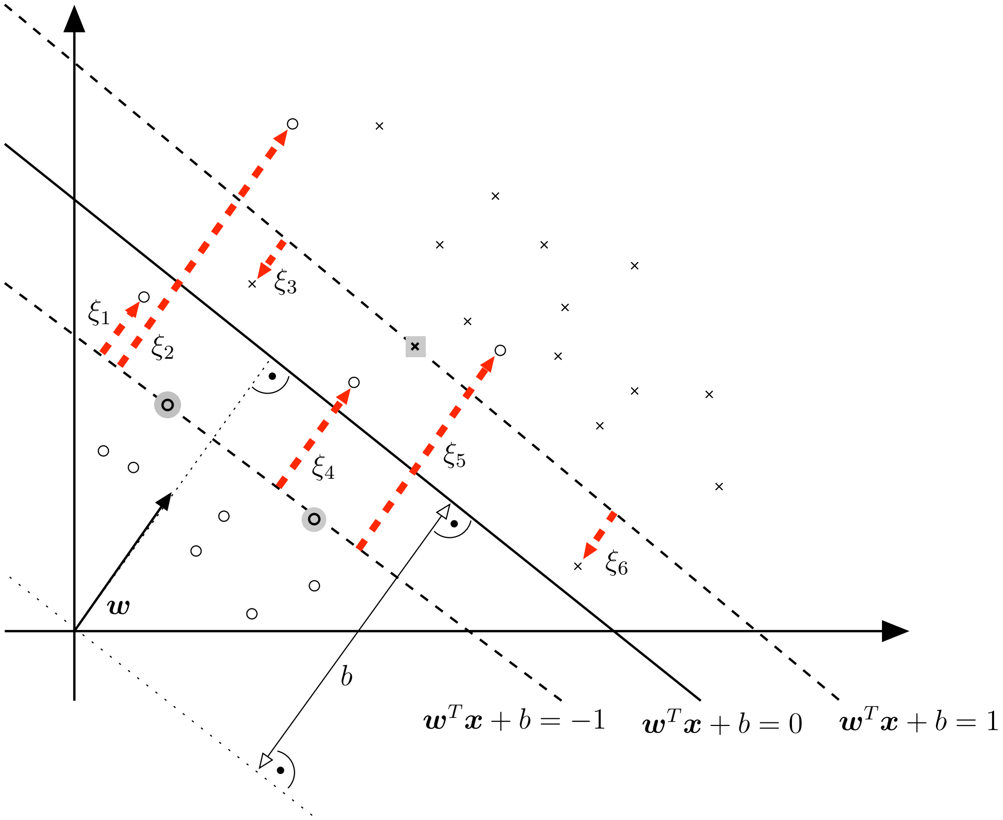

ξ_1
ξ_2
ξ_3
ξ_4
ξ_5
ξ_6
$\boldsymbol{w}$
b
$\boldsymbol{w}^T\boldsymbol{x} + b = -1$
$\boldsymbol{w}^T\boldsymbol{x} + b = 0$
$\boldsymbol{w}^T\boldsymbol{x} + b = 1$

Figure 2.3: Noisy data is handled by slack variables ξ_i which allows some data points to lie within the margin or even on the "wrong" side of the hyperplane.

This is illustrated in Figure 2.4. The key to the success of this approach lies within eqn. (2.12). Interestingly, the data $\vec{x}$ and $\vec{x}_i$ enter eqn. (2.12) only by their scalar product. As the function Φ might be expensive to compute or even map to an infinitely dimensional space, one is more interested in functions which perform both, the mapping and scalar product computation. Such a function is called a *kernel* k:

$$k(\vec{x}, \vec{x}_i) = \Phi(\vec{x})^T \Phi(\vec{x}_i) \tag{2.18}$$

To assure that there exists a space in which the kernel computes the scalar product implicitly, one uses kernels that satisfy the Mercer condition - so called *Mercer Kernels*. A popular choices for this kernel is the RBF kernel:

$$k(\vec{x}, \vec{x}_i) = \exp(-\frac{||\vec{x} - \vec{x}_i||^2}{2\sigma^2}) \ , \ \sigma \in \mathbb{R}^+ \tag{2.19}$$

For other valid choices and how to compose new kernel functions we refer to Cristianini and Shawe-Taylor (2000) and Chapelle *et al.* (1999).

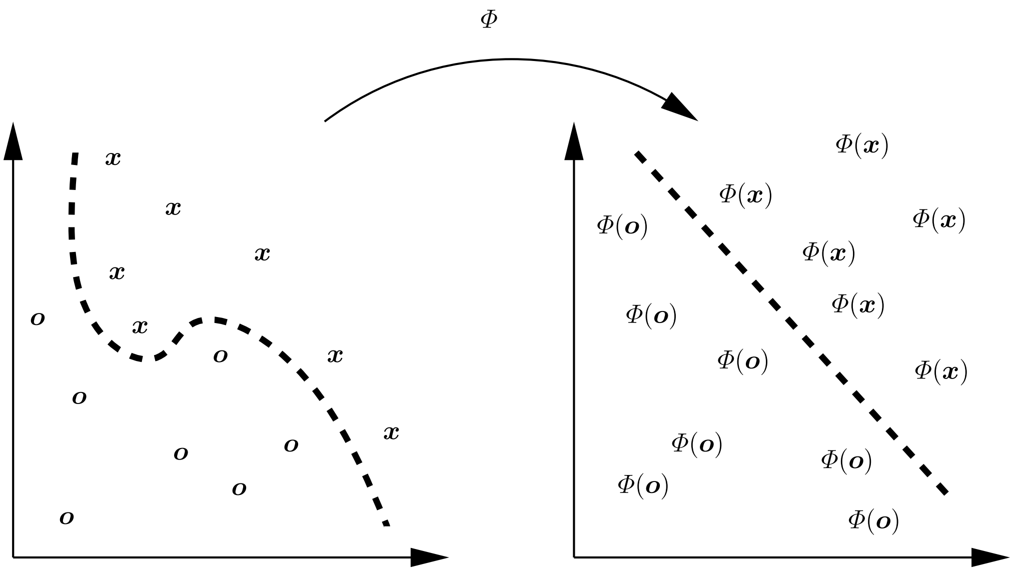


Figure 2.4: Illustration of a non-linear mapping Φ which makes the classes in the data linearly separable.

2.2.3 Topic Models

In terms of generative modeling, we build on the success of topic models (e.g. Hofmann (2001); Blei *et al.* (2003b); Griffiths and Steyvers (2004)). They have gained increasing attention in computer vision ranging from unsupervised category discovery (Sivic *et al.*, 2005; Li and Perona, 2005; Bosch *et al.*, 2006), over classification (Quelhas *et al.*, 2005; Larlus and Jurie, 2006) to detection (Sudderth *et al.*, 2005; Fergus *et al.*, 2005a; Bissacco *et al.*, 2008). Often local feature representations are employed (Sivic *et al.*, 2005; Larlus and Jurie, 2006) that neglect the spatial layout with a few exceptions such as (Fergus *et al.*, 2005a; Sudderth *et al.*, 2005; Bissacco *et al.*, 2008).

We employ probabilistic topic models as described in (Hofmann, 2001; Blei *et al.*, 2003b; Griffiths and Steyvers, 2004) which were originally motivated in the context of text analysis. As it is common habit we adopt the terminology of this domain. In the following, a document d refers to a sequence of words $(w_1, w_2, \ldots, w_{N_d})$, where each w_i is one word occurrence. The underlying idea of these models is to regard each document as a mixture of topics. This means that each word w_i of the total N_d words in document d is generated by first sampling a topic z_i from a multinomial topic distribution $P(z)$ and then sampling a word from a multinomial topic-word distribution $P(w|z)$. Therefore the word probabilities for the combined model are:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{2.20}$$

where T is the number of topics and $P(w_i|z_i = j)$ as well as $P(z_i = j)$ are unob-

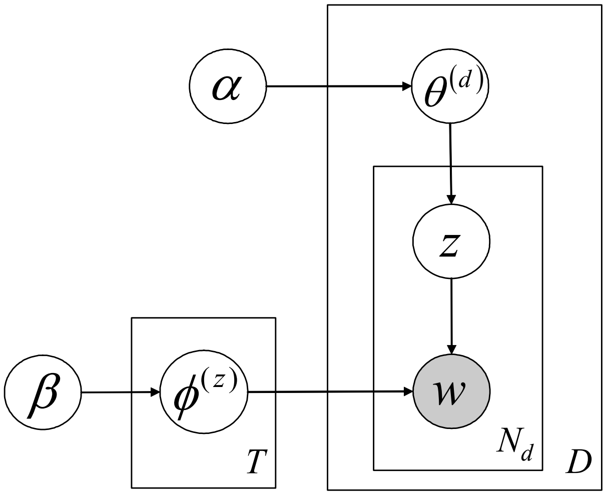


Figure 2.5: LDA model as formulated by Griffiths and Steyvers (2004).

served. According to the notation of Griffiths and Steyvers (2004), we will abbreviate

$\theta^{(d)}$: topic distribution $P(z)$ for document d
$\phi^{(j)}$: topic-word distribution $P(w_i|z = j)$ for topic j

The particular topic models differ on the one hand in which additional hyperparameters/priors they introduce and on the other hand in how inference and parameter estimation is performed. We will discuss the *Latent Dirichlet Allocation* model of Blei *et al.* (2003b) in some more detail focusing on the version presented in Griffiths and Steyvers (2004) that uses Gibbs sampling for inference and estimation. The graphical representation of this model is depicted in Figure 2.5. It visualizes the process that generates a total of D documents d, where each document has N_d words. Above we already described how each word w_i of a particular document is generated. In the full model, there are 2 additional hyperparameters, α and β, which place symmetric dirichlet priors on the topic distribution of each document $\theta^{(d)}$ and the topic-word distributions $\phi^{(j)}$ respectively. As the setting for α and β is common to all documents, these act as forces that impose global tendencies on these distributions. Intuitively, the prior α for the topic distribution θ favors co-activation (sharing) of multiple topics for each document for values larger than 1, whereas smaller values result in sparser topic distribution - ultimately having single topics explaining whole documents (clustering). Consequently, the sparseness of the topic-word distribution $\phi^{(j)}$ is affected by this choice. The second parameter β, has a direct smoothing effect on the topic distributions.

For more details on the models, inference and estimation, we refer to Blei *et al.* (2003b) and Steyvers and Griffiths (2007). The idea behind the employed Gibbs sampling procedure is that all topic assignments z_i are initialized (typically randomly) and then iteratively updated in a random order. To perform such a single update, a topic is drawn from the conditional distribution $P(z_i|\Omega \setminus z_i)$ and assigned to z_i, where $\Omega \setminus z_i$ denotes all observed and unobserved variables but z_i. This is repeated for a fixed number of iterations.

2.3 Inspiration by Previous Work for the Contributions of this Thesis

In this section we describe how the main contribution of this thesis were inspired by previous work.

2.3.1 Generative/Discriminative Hybrid Model for Detection

Inspired by the different properties of generative and discriminative approaches as described in Section 2.1.1 and work from the machine learning community (Jaakkola and Haussler, 1999), we present a hybrid approach in Chapter 3. We use the Implicit Shape Model (ISM) (Leibe *et al.*, 2008) as the generative part which is know to obtain high recall and generalize well from few training instances. But these benefits come at the price of over simplified independence assumptions of a star-structured model. This is why we decided for the kernel proposed in Wallraven *et al.* (2003) as the discriminant counterpart as it allows to verify the hypotheses with a stronger shape models that ensures global consistency.

2.3.2 Weakly Supervised Learning by Discovery of Reoccurring Patterns

Typically, highly supervised methods like the ISM that use pixel-level annotations lead to superior performance in comparison to unsupervised (Sivic *et al.*, 2005) or weakly supervised (Fergus *et al.*, 2003) methods. This motivated us to formulate an approach that bridges the gap between low annotation effort and highly supervised performance level. Chapter 4 presents an approach that first recovers object annotations as reoccurring patterns in a weakly supervised setting from data, which allows then training of fully supervised models. The metric we use to retrieve those reoccurring patterns also has its origin in the ISM. The notion of Scale-Invariant Patterns which we are going to introduce in Chapter 4 is set up in a way, that the scalar product between two patterns mimics the ISM voting scheme described in Section 2.2.1.

2.3.3 Integrating Different Levels of Supervision in a Cross-Modal Setting

The before mentioned approaches mostly use a single type of supervision. As part of the work in Chapter 5 is motivated by a tutor-driven learning scenario there are two main reasons which motivated us to use combine different levels of supervision in one approach. First, tutor-based interactions are very time consuming so that we want to tap into any kind of information that is available to the system. In our case, this means exploiting unsupervised clustering as well as uncertain information in addition to the direct supervision by the tutor. Second, supervision is needed to overcome the inherent limitations of purely unsupervised approaches. While fully unsupervised approaches like Sivic *et al.* (2005) have shown impressive performance, there is a limit to what can be inferred by just analyzing the data.

2.3.4 Generative Decompositions of Visual Categories

As outlined in Section 2.1.2 the diversity of representation paradigms in the literature motivated us to start in Chapter 6 from a rather low level description of the image based on oriented gradient histograms (Dalal and Triggs, 2005). Instead of hand-crafting the features we aim for learning a representation from data. Also from the perspective of topic modeling, a change towards a dense representation as also promoted in Bissacco *et al.* (2008) is beneficial in three ways. First, one is less likely to miss important information in contrast to a sparse local feature representation which relies on an interest point detector. Second the employed dense representations lend themselves to visual inspection which turned out to be extremely helpful in order to gain more model introspection. Third, the dense representation in combination with the probabilistic topic model eliminates the need of codebook generation and matching, which can introduce additional noise to the system. In particular, over-counting evidence due to overlapping features as it can occur for local features is fully eliminated. Finally, approaches like (Quelhas *et al.*, 2005) inspired us to use the topic representation as an intermediate result. This is in contrast to approaches like Sivic *et al.* (2005) where categories are discovered as single topics. Torralba *et al.* (2007) inspired us to learn a shared representation across all classes to improve scalability.

3

Integrated Representative/Discriminative Approach

This chapter presents a method for object category detection which integrates a generative model with a discriminative classifier. For each object category, we generate an appearance codebook, which becomes a common vocabulary for the generative and discriminative methods. Given a query image, the generative part of the algorithm finds a set of hypotheses and estimates their support in location and scale. Then, the discriminative part verifies each hypothesis on the same codebook matches. The new algorithm exploits the strengths of both original methods, minimizing their weaknesses. Experiments on several databases show that our new approach performs better than its building blocks taken separately. Moreover, experiments on two challenging multi-scale databases show that our new algorithm outperforms previously reported results.

More specifically, we combine the Implicit Shape Model (ISM, Leibe *et al.* (2004)) based on a codebook representation (which can be seen as a non-parametric probabilistic model of the appearance of object categories) with an SVM using Local Kernels (LK, Wallraven *et al.* (2003)), which has proven effective for object categorization (Nilsback and Caputo, 2004). The idea to use a generative model inside a kernel function has been proposed before (Jaakkola and Haussler, 1999; Jebara *et al.*, 2004; Vasconcelos *et al.*, 2004; Tsuda *et al.*, 2002), and it has been applied to visual recognition tasks like object identification (Vasconcelos *et al.*, 2004).

The rest of the chapter is organized as follows: after a brief outline of the building blocks (Section 3.1), we introduce the new approach, describing in detail how it integrates ISM and LK and discussing its advantages with respect to the two previous methods (Section 3.2). Section 3.3 reports experiments benchmarking our new method with its building blocks, on several databases of increasing difficulty.

3.1 Previous Approaches

Our approach is motivated by two recent advances in object detection and discriminant classification.

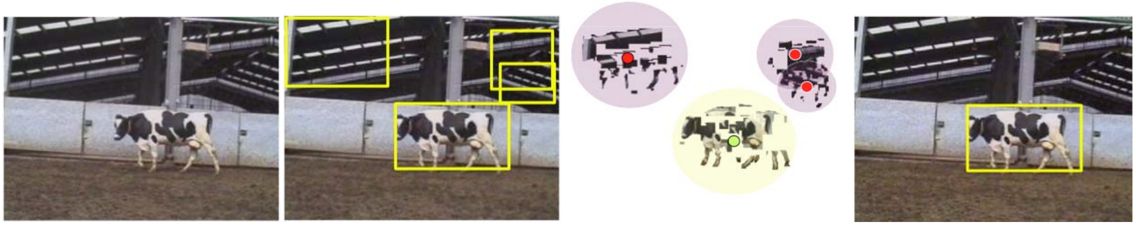
Object Detection with Implicit Shape Models. As detailed in Section 2.2.1, Implicit Shape Models (ISMs) (Leibe *et al.*, 2008) are unique in that they address object category detection and top-down segmentation at the same time. They proceed by first collecting the evidence from local features in a probabilistic Hough voting procedure to determine possible object locations and scales. For each such hypothesis, they then go back to the image to determine on a per-pixel level where its support came from, thus effectively segmenting the object from the background. The segmentation information can then in turn be used to improve the accuracy of the detection and resolve ambiguities between overlapping hypotheses (Leibe *et al.*, 2004). As a result of this iterative process, ISMs have been shown to yield good object detection results and considerable robustness to partial occlusion.

The ISM approach provides a flexible representation of the target category. Since each image patch votes for the object center independently of the other patches, the resulting model can interpolate between local parts seen on different training objects. As a result, it can adapt well to novel objects of the target category and typically achieves high recall. However, as a price for this flexibility, it cannot reject false positives as accurately as a discriminative model.

SVM Classification with Local Kernels. Most current object category detection systems are based on local features in order to reduce the influence of intra-class variations, noise, and occlusion (Fergus *et al.*, 2003; Agarwal *et al.*, 2004; Viola *et al.*, 2003; Viola and Jones, 2004; Lowe, 2004; Leibe *et al.*, 2004). Support Vector Machines (SVMs), on the other hand, have shown impressive learning and recognition performance (Pontil and Verri, 1998; Papageorgiou and Poggio, 2000; Heisele *et al.*, 2001). As the SVMs' machinery requires the computation of scalar products on the feature vectors, Wallraven *et al.* (2003) introduced a local kernel which formulates the feature matching step as part of the kernel itself. Despite the claim in Wallraven *et al.* (2003), this family of kernels is not a Mercer kernel (Boughorbel *et al.*, 2004). Still, it can be shown that it statistically approximates a Mercer kernel in a way that makes it a suitable kernel for visual applications. On the basis of this finding, and of its reported effectiveness for object categorization (Nilsback and Caputo, 2004), we will use this family of kernels in this chapter.

Given two sets of local feature L_h and L_k, these local kernels are defined as (Wallraven *et al.*, 2003):

$$K(L_h, L_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots,n_k} \left\{ K_l(L_h^{j_h}, L_k^{j_k}) \right\}, \tag{3.1}$$



(a) original image (b) detected hypotheses (c) SVM input (d) accepted hypotheses

Figure 3.1: Stages of the integrated approach. (a) original image; (b) hypotheses detected by the representative ISM; (c) input to the SVM stage; (d) verified hypotheses.

where the local feature similarity kernel K_l consists of an appearance part K_a and a position constraint K_p

$$K_l(L_h^a, L_k^b) = K_a(L_h^a, L_k^b)\ K_p(pos(L_h^a), pos(L_k^b)). \tag{3.2}$$

Various options have been given for the selection of K_a and K_p Wallraven *et al.* (2003), including the following choice

$$K_a(x,y) = exp\left\{-\gamma\left(1 - \frac{\langle x-\mu_x | y-\mu_y\rangle}{||x-\mu_x||||y-\mu_y||}\right)\right\} \tag{3.3}$$

$$K_p(\lambda_x,\lambda_y) = \exp\left(-\frac{(\lambda_x-\lambda_y)^2}{2\sigma^2}\right), \tag{3.4}$$

where σ is a strictness parameter for the position constraint.

As shown by Wallraven *et al.* (2003) and Nilsback and Caputo (2004), Local SVMs can discriminate well between different object categories. However, they contain no localization component and require accurate initialization in position and scale. In the literature, the standard solution to this problem is to perform an exhaustive search over all possible object positions and scales Papageorgiou and Poggio (2000); Schneiderman and Kanade (2000); Heisele *et al.* (2001); Viola and Jones (2004); Agarwal *et al.* (2004); Torralba *et al.* (2004). However, this exhaustive search imposes severe constraints, both on the detector's computational complexity and on its discriminance, since a large number of potential false positives need to be excluded. In this chapter, we present a different solution to this problem by integrating Local SVMs with the ISM approach.

3.2 Integrated Approach

The main contribution of this chapter is to integrate both approaches into a consistent framework (visualized in Fig. 3.1). Applied to a novel test image (Fig. 3.1(a)), the representative ISM is first used to find a set of promising hypotheses (Fig. 3.1(b)) and estimate their support in both location and scale (Fig. 3.1(c)). For each of those hypotheses, the more exact discriminative model is then applied in order to verify them and filter out false positives (Fig. 3.1(d)). By using the same internal representation, namely the appearance codebooks, those two approaches are tightly integrated. The ISM uses these appearance codebooks to generate hypotheses which are visually consistent and which follow a weak spatial model. The discriminative model on the other hand uses the same appearance codebooks to find visually discriminant information for object classes and also to add a stronger spatial model effectively extracting discriminant spatial codebook distributions. We thus combine the capabilities of both models in an advantageous manner.

3.2.1 Generation of an Appearance Codebook

As a common representation, we generate a category-specific appearance codebook, as described in Leibe *et al.* (2004). For this, we apply a scale-invariant DoG interest point operator Lowe (2004) to all training images and extract image patches with a radius of 3σ of the detected scale. All extracted patches are then rescaled to a uniform size (in our case 25×25 pixels) and grouped using an agglomerative clustering scheme. The resulting clusters form a compact representation of local object structure. In the following, we keep only the cluster centers $C = (\vec{c}_1, \ldots, \vec{c}_R)$ as codebook entries.

3.2.2 Representation in Codebook Coordinates

The result of the ISM stage (see Section 2.2.1) is a set of object hypotheses $h = (o_n, \lambda)$, together with their support in the image (Figure 3.1(c)). This support consists of a list of local features that contributed to the hypothesis and their corresponding codebook matches. In order to interpret this information in the SVM framework, we first have to adapt the kernel formulation to our codebook representation.

The key idea is that the scalar product $\langle \vec{x}, \vec{y} \rangle$ used in the SVM Kernel can be expressed in terms of a codebook matching problem. For this, we project both $\vec{x}$ and $\vec{y}$ into the affine space spanned by the codebook entries $\vec{c}_i$ as basis vectors. With $\vec{x} = \sum_i a_i \vec{c}_i$ and $\vec{y} = \sum_j b_j \vec{c}_j$ the scalar product can be written as

$$\langle \vec{x}, \vec{y} \rangle = \sum_i \sum_j a_i \langle \vec{c}_i, \vec{c}_j \rangle \, b_j. \tag{3.5}$$

This formulation has two advantages. One is its computational efficiency – both the intra-codebook similarity matrix $\langle \vec{c}_i, \vec{c}_j \rangle$ and the support vector coefficients b_j can be precomputed. Only the image-feature coefficients a_i need to be calculated during recognition. The second advantage is that the data is now expressed in a common format and partial results can be reused by both stages.

Remains the problem how to select the coefficients a_i and b_j. The smallest reconstruction error would be obtained by a least-squares solution, but this solution is typically not sparse. In order to arrive at a sparse representation, we propose to consider again only the closest-matching codebook entries $C^*_{\vec{x}} = \{\vec{c}^*_i | sim(\vec{c}^*_i, \vec{x}) \geq \theta\}$ and approximate the vectors $\vec{x}$ and $\vec{y}$ by the mean of those "activated" codebooks. Thus, with $n = |C^*_{\vec{x}}|$, $m = |C^*_{\vec{y}}|$, we arrive at

$$\begin{aligned} \langle \vec{x}, \vec{y} \rangle \approx \langle \vec{\mu}_x, \vec{\mu}_y \rangle &= < \frac{1}{n} \sum_{i=1}^{n} \vec{c}^*_i, \frac{1}{m} \sum_{j=1}^{m} \vec{c}^*_j > && (3.6) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{n} \left\langle \vec{c}^*_i, \vec{c}^*_j \right\rangle \frac{1}{m}. && (3.7) \end{aligned}$$

This approximation is justified under the assumption that the codebook entries sufficiently cover the relevant "object" region of the appearance space. We have verified the validity of this assumption in a series of control experiments. The results indicate that the difference in reconstruction error between the least-squares solution and our sparse approximation is only modest and subsumes to an average error of approximately one gray level per pixel on a reconstructed patch. As a result, we get a problem-specific representation which expresses the data in a common vocabulary and is used throughout both stages of our approach. In particular, this representation allows us to reuse the results of the initial codebook matching stage for the SVM model.

3.2.3 SVM Verification with Local Kernels

Let $X = \{(x_1, \lambda_1), \ldots, (x_N, \lambda_N)\}$ be a set of local features (with appearance and relative location) supporting hypothesis h, and $A = \{A_1, \ldots, A_N\}, A_i = (a_1, \ldots, a_R)$ be their corresponding codebook matches. The ISM procedure guarantees that each feature in the supporting set is consistent in appearance and location with at least one training example. However, as only local consistency is enforced, this reference example may be a different one for each feature. In the next step, we therefore want to verify that the global feature configuration is also consistent.

Figure 3.2 visualizes the chosen verification procedure. In the remainder of this section, we will define the Local Kernel in a way that each support vector corresponds to a distinct configuration of local features. When evaluating a hypothesis, it is successively compared to each support vector. For each such match, correspondences are established between visually similar features occurring in the same relative locations (with a small tolerance σ), and the quality of the resulting global configuration fit is measured. Thus, the kernel enforces strong spatial constraints to verify the hypothesis.

This is done as follows. Let $Y = \{(y_1, \lambda_1), \ldots, (y_M, \lambda_M)\}$ be the features observed on a training image with corresponding codebook activations $B = \{B_1, \ldots, B_M\}, B_j = (b_1, \ldots, b_R)$. In order to compare the feature configurations of X and Y, we first try

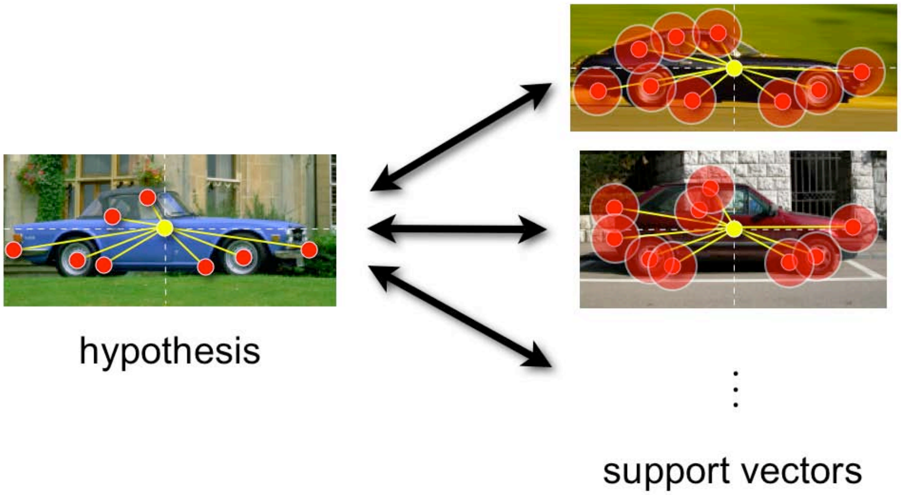


Figure 3.2: A look inside the LK verification stage. Each support vector specifies a configuration of local features, corresponding to a particular training example. When evaluating a hypothesis, the kernel K first searches for the k best feature correspondences, considering both appearance and relative position, and uses them to judge the quality of the match.

to find a set of correspondences between their features. For each pair of features $(\vec{x}, \lambda_x)$ and $(\vec{y}, \lambda_y)$, the quality of a match is expressed by the local similarity kernel K_l:

$$K_l((\vec{x}, \lambda_x), (\vec{y}, \lambda_y)) = K_a(\vec{x}, \vec{y})\; K_p(\lambda_x, \lambda_y), \tag{3.8}$$

where K_a is measuring the appearance similarity and K_p is imposing a position constraint in the manner of a penalty function. For K_a and K_p we stick to the choices made in Wallraven *et al.* (2003), but replace the correlation coefficient by the approximation from eq. (3.5):

$$\begin{aligned} K_a(\vec{x}, \vec{y}) &= \exp\left(-\gamma\left(1 - \langle \vec{x}, \vec{y} \rangle\right)\right) && (3.9) \\ &\approx \exp(-\gamma(1 - \sum_i \sum_j a_i \langle \vec{c}_i, \vec{c}_j \rangle\, b_j)) \\ K_p(\lambda_x, \lambda_y) &= \exp\left(-\frac{(\lambda_x - \lambda_y)^2}{2\sigma^2}\right). && (3.10) \end{aligned}$$

In order to allow for some flexibility in the part arrangement, we do not enforce complete correspondence, but only match a subset of the features by searching for the k best correspondences. This is done using a greedy selection strategy on the feature similarity matrix $K_l(X, Y)$. Let $\Phi \in \pi_1^N$, $\Psi \in \pi_1^M$ be permutations of the local features to reflect this greedy assignment. According to Wallraven *et al.* (2003), the corresponding Local Kernel is then defined as

$$K(X, Y) = \frac{1}{k} \max_{\Phi, \Psi} \sum_{j=1}^{k} K_l\left((\vec{x}_{\Phi(j)}, \lambda_{x,\Phi(j)}), (\vec{y}_{\Psi(j)}, \lambda_{y,\Psi(j)})\right). \tag{3.11}$$

Note that the resulting kernel does not need to consider the original features anymore, but only operates on the codebook matches passed from the previous stage. It thus requires very little computation and imposes only a small overhead on the total execution time. In all experiments presented in this paper, we set k to 50 and determine the remaining parameters using cross-validation on the training set. Although it might be tempting to set $k = \max(N, M)$ and then normalize the kernel by dividing with k, this has shown to lead to a kernel matrix that is not positive semi-definite. The SVM training on those kernels lead to extremely poor results.

3.2.4 Discussion

It is important to emphasize that through the integration, the SVM stage is solving a simpler problem than the previous LK approach. Not only is it initialized with an estimate of the object position and scale, but it directly obtains also the supporting image features as input. It can thus optimize its decision surface on the failure cases of the ISM stage and learn a stronger discriminative model. In addition, the discriminative model makes it possible to achieve a better separation from background constellations, whose complex distributions are notoriously hard to express in a probabilistic framework.

The matching to a common codebook enables both stages to make use of “across-instance” learning which is essential when dealing with limited training set sizes. In addition, the Local Kernel stage complements the ISM’s weak spatial model with stronger spatial constraints.

As a side benefit, the output confidence of the SVM stage (i.e. the distance to the hyperplane) becomes comparable for different object categories. This is the case because the Local Kernel bases its computation on a fixed number of k correspondences.

3.3 Experiments

In this section, we show that our new approach benefits from the integrated representative and discriminative representation (in the following termed IRD). We present our results in three steps. After presenting the data, we first compare our new approach to the original ISM and LK approaches in Section 3.3.2. Section 3.3.3 then reports results on a multi-category detection/discrimination task. Finally, we evaluate our approach on two difficult data sets and the PASCAL VOC challenge (Everingham *et al.*, 2005b) containing large scale changes and partial occlusion

3.3.1 Data

In order to evaluate our approach, we apply it to a test set containing objects of four categories, namely cars, cows, horses, and motorbikes. The pairs cars/motobikes and cows/horses were especially chosen to measure cross-category confusions, since they

	LK	ISM	IRD
car	61.0 %	94.7 %	**99.4%**
cow	95.3 %	96.1 %	**97.1%**
horse	77.8 %	88.5 %	**88.5%**
motorbike	87.6 %	93.8 %	**96.5%**

Table 3.1: Equal error rate performances achieved by the Local Kernel (LK), Implicit Shape Model (ISM), and our integrated approach (IRD) on present/absent tasks.

share similar visual features. The data is mostly taken from the PASCAL database collection (Everingham *et al.*, 2005a). For cars we use the UIUC single-scale test set; for motorbikes the CalTech motorbike set (with the same training/test split as in Fergus *et al.* (2003)); and for cows the TUD cow database (supplemented with 557 test images). For the background set, we use 450 CalTech background images. The horse images are taken from the Weizmann horse database (Borenstein and Ullman, 2002) and split into 79 training and 164 test images. This is the first time detection results are reported on this database, as it was previously only used for segmentation tasks.

3.3.2 Comparison with Original Approaches

We start the experimental part with a comparison of our new IRD approach with the approaches it originates from – namely the Local Kernels and the ISM. To provide a fair comparison with the LK approach, which is not designed to be scale invariant or perform a detection task in the first place, we report results of object present/absent experiments. The test is performed on images of each category vs. 450 novel background images.

Table 3.1 summarizes the equal error rate (EER) performances for this experiment. As can be seen from the table, the integrated approach achieves superior performance compared to its building blocks.

3.3.3 Multi-Category Discrimination

Detection Task and Evaluation. Our main experiments are performed on a detection task, where the detector has to localize image regions in which an instance of the category of interest is present. For evaluating the car detections, we use exactly the same acceptance criterion as in Agarwal *et al.* (2004). However, as this criterion is only well-defined for fixed-size bounding boxes (and thus not directly applicable to the cow and horse categories), we apply an extended criterion for the other three categories. We inscribe an ellipse in the ground truth bounding box and measure the distance d_r between the bounding box centers relative to the ellipse's radius at the corresponding angle. A hypothesis is accepted if $d_r \leq 0.5$ and the ground truth and hypothesis bounding boxes cover one another by at least 50%. In

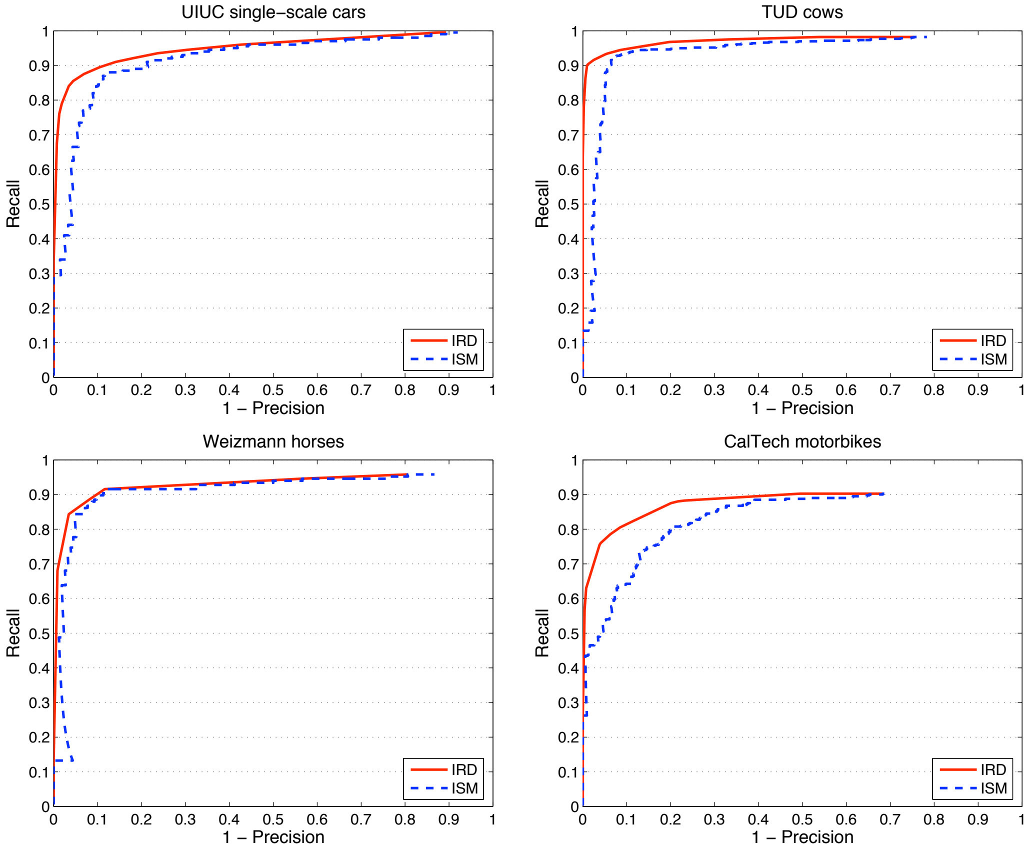


Figure 3.3: Recall-Precision curves for the car, cow, horse and motorbike model on a detection task.

accordance to Agarwal *et al.* (2004), only one hypothesis per object is accepted as correct – any additional hypothesis on the same object is counted as false positive.

Detection Results. Figure 3.3 shows the results of this evaluation in the form of Recall-Precision curves (RPCs). To vary the strictness of the local kernel SVM in our new IRD approach without retraining, we used the distance to the decision boundary as a confidence measure. Although the ISM by itself performs already quite well on all four categories, our new IRD approach improves the EER performance for cars from 87.6% to 88.6%, for cows from 92.5% to 93.2% and for motorbikes from 80.0% to 84.0%. For horses, the performance stays at the same level. Besides the gain in EER performance, cars, cows and motorbikes profit from the added discriminance in terms of increased precision of the final detector, which shifts the precision-recall curves to the left. Especially the relatively rigid car and motorbike categories profit from the stronger structured constraint of the local kernel. Figure 3.5 displays some example detections of our approach which illustrate the generalization capabilities over large intra-category variations, including different articulations, and its robustness to partial occlusion.

ISM **IRD**	car #170	cow #557	horse #164	motorbike #400
car	-	0.07 **0.00**	0.02 **0.01**	0.18 **0.03**
cow	1.00 **0.49**	-	0.18 **0.11**	1.05 **0.05**
horse	0.71 **0.16**	0.53 **0.08**	-	0.68 **0.05**
motorbike	1.07 **0.08**	0.29 **0.09**	0.22 **0.00**	-

Figure 3.4: Cross-category confusions (false positives per test image) for the ISM and our new IRD approach on a detection task.

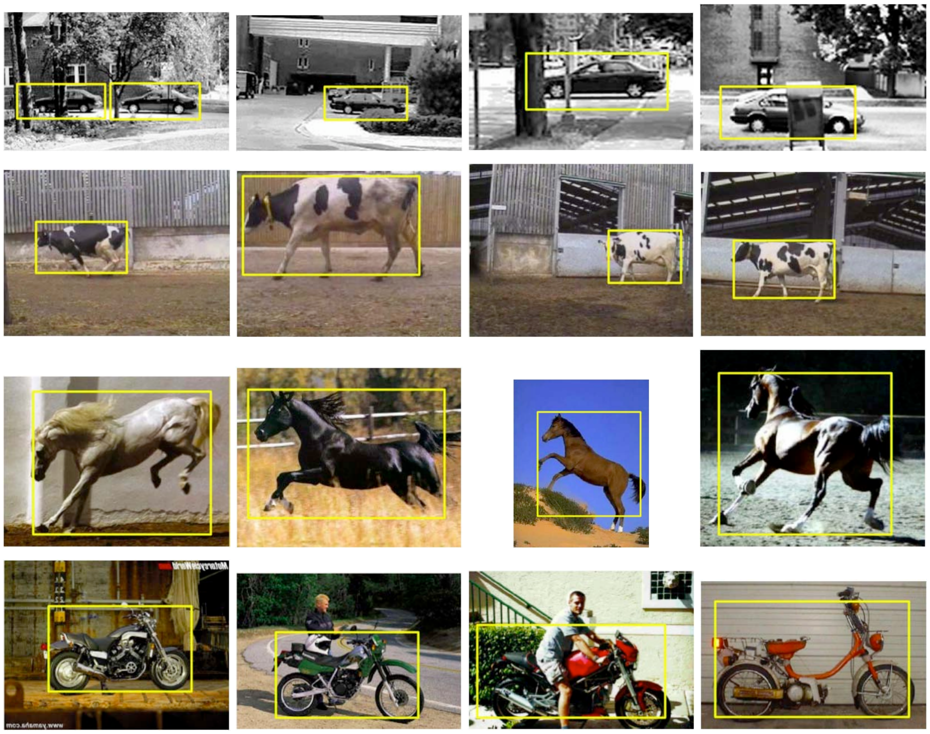


Figure 3.5: Example detections on the car, cow, horse, and motorbike test sets.

Discrimination Results. Given these detectors operating at their equal error rate, we now investigate the produced confusions. Table 3.4 displays the false detections each of the detectors produces per image on all four object categories. The left number reports the false positives detected by the ISM. It can be seen that the ISM performs well for the car model, but still produces a relatively large number of false positives for the other categories. The larger number of confusions on the car images can be explained by the fact that those images are about twice as large as the other images. The right number reports the false positives detected by our new IRD approach processing all detectors in parallel and acting as a single unified detector. Ambiguous detections are eliminated using the local SVM output as a confidence measure. We can observe a drastic reduction of false positives down to (or even below) the 0.1 level for almost all combinations. In particular, these results

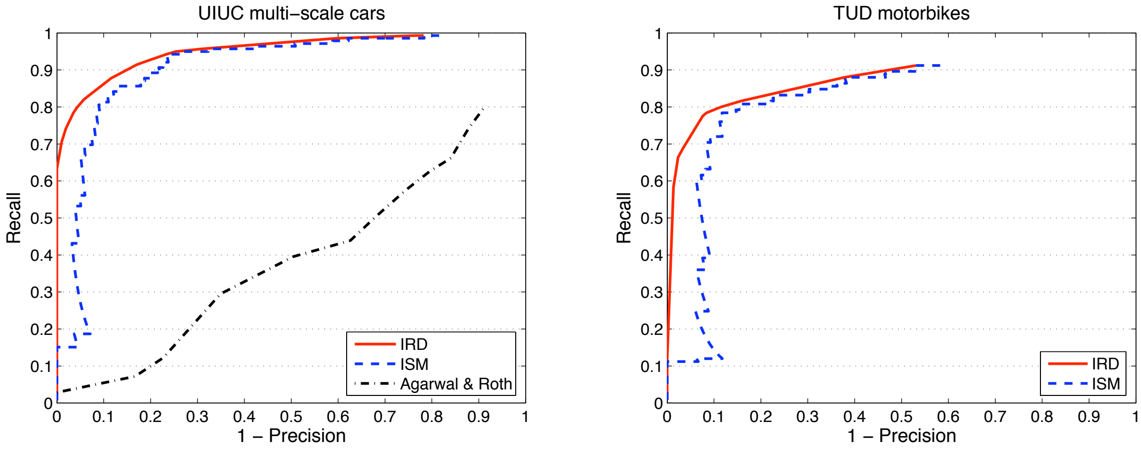


Figure 3.6: Precision-Recall curves for the difficult multi-scale databases of cars and motorbikes. Our new IRD approach clearly outperforms the state-of-the-art on the UIUC multi-scale car database.

show that in our new IRD approach the SVM output is well suited as a confidence measure for comparing hypotheses across categories.

3.3.4 Discriminant Category Detection

In this section, we evaluate our approach on two more challenging databases that include large scale changes and significant partial occlusion. We use the UIUC multi-scale cars (Agarwal *et al.*, 2004) and the TUD motorbikes. For the multi-scale cars, we again use the acceptance criterion from Agarwal *et al.* (2004); for the motorbikes we use the criterion described in Section 3.3.3 for the reasons given there. Figure 3.6 shows the result of this evaluation. The black line corresponds to the performance reported by Agarwal *et al.* (2004), with an EER performance of about 45%. In contrast, our IRD approach achieves 87.8% EER – an improvement of over 40%! Interestingly, our method obtains up to 64% recall before generating any false positives. On the motorbike test set, our approach achieves an EER performance of 81%. Compared to ISM, there is a consistent improvement in precision. The difficulty of the task is illustrated by Figure 3.7, which shows example detections of our IRD approach documenting the performance under occlusion, extreme illumination conditions and large scale changes.

3.3.5 Results on the PASCAL VOC Challenge 2005

We entered the PASCAL challenge with results on the tasks No. 1, 2, 5, 6 - namely detection and classification on test set 1 and 2, where the model is trained on the provided training and validation sets. We submit results on the categories car and motorbike obtained with the standard ISM approach as well as our new IRD

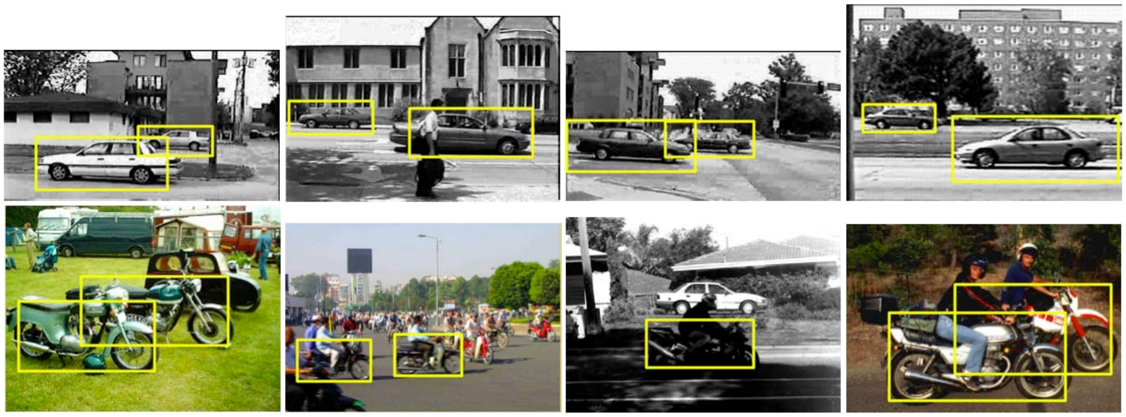


Figure 3.7: Example detections of our new IRD approach on the difficult multi-scale UIUC car database and the multi-scale TUD motorbike test set.

approach.

Training. The ISMs are trained on the following subsets of the training/validation sets:

- 55 car images consisting of
  - 26 images of the TU Darmstadt database
  - 29 images of the TU Graz database
- 153 motorbike images of the CalTech database

Up to now, our approaches have only been evaluated on single viewpoints. In order to stay consistent with those experiments, we only selected side views from the training set. .

Again, SVM validation stage is trained on detections and false alarms of the ISM on the whole training set for cars and motorbikes.

Test. All experiments were performed on the test sets exactly as specified in the PASCAL challenge. For computational reasons, the test images were rescaled to a uniform width of 400 pixels, as otherwise the number of local features cannot be handled well by the non-parametric approach we are taking in the generative detection stage.

We report results on both the object detection and the present/absent classification task. Detection performance is evaluated using the hypothesis bounding boxes returned by the ISM approach. For the classification task, an object-present decision is taken if at least one hypothesis is detected in an image.

Since our IRD approach allows for an additional precision/recall trade-off, we explored different trade-offs for our submission to the challenge – one for optimal equal

database	percentage	#cars
CalTech	6%	21
TU Darmstadt	7%	24
TU Graz	45%	152
UIUC	42%	144

Table 3.2: Analysis of the image sources for the car test set 1. Only the TUD and UIUC databases contain exclusively side views. Altogether, side views take up only about 55% of the test cases.

error rate (EER) performance and one for optimized precision (labeled “ISM+SVM v2” in the plots).

Results on Motorbikes Figure 3.8 shows our results for the motorbike test sets. With 89.5% and 85.5% EER performance for detection and classification, respectively, our methods achieve good performace on test set 1. On test set 2, the performance drops to 50% and 68% . This can be explained by the problem of detecting/classifying multiple view points. As mentioned above, our approach is only trained on side views, which comprise only a small portion of the test images. Examples for the detection task on test set 1 and 2 are shown in Figure 3.11. Side by side, the predicted bounding box and the inferred segmentation mask is shown. The detection of the motorbike at the beach exposes an interesting artifact. As the model was only trained on motorbikes facing to the right, this left facing motorbike gets interpreted as a right facing one, as can be seen from the segmentation mask. While this works out well in this particular case for the motorbike category, it is not trivial how to deal with these kind of problems in a more principled way. For a possible approach to extend the ISM towards multi-view detection and out-of-plane rotations, we refer the reader to Thomas *et al.* (2006).

Results for Cars Our results for the car test sets are shown in Figure 3.9. As can be seen from those plots, the achieved recall is significantly lower than for the motorbike test sets (with 53% for detection and 64% for classification on test set 1). The reason for this is again that our approach is only trained on side views. As Table 3.2 shows, only about 55% of the test cases consist of side views, which forms an upper bound on our detector’s performance. However, as the “ISM+SVM v2” curve demonstrates, our approach succeeds to find a large percentage of the existing side views with high precision. Example detection are depicted in Figure 3.13.

The same observation holds for test set 2, where 10% of the cars are detected with high precision, which again corresponds to a large percentage of the existing side views in this test set. The composition of the test sets is also reflected in our approach’s performance for the classification task, where it achieves 66% EER performance on test set 2.

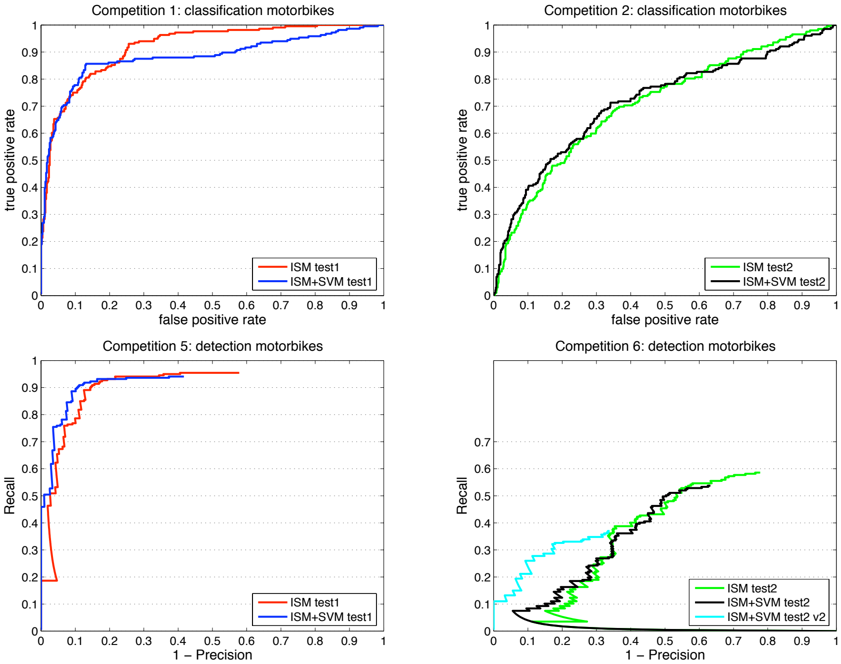


Figure 3.8: Results on test sets 1 (left) and 2 (right) for motorbikes.The top row shows the performance for the object present/absent classification task. The bottom row contains our results for the object detection task.

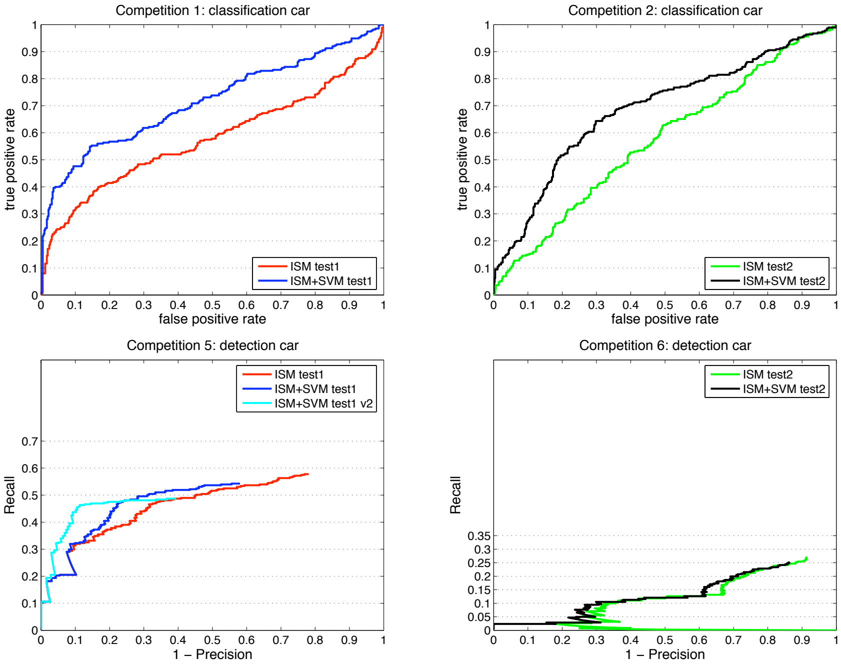


Figure 3.9: Results on test sets 1 (left) and 2 (right) for cars.The top row shows the performance for the object present/absent classification task. The bottom row contains our results for the object detection task.

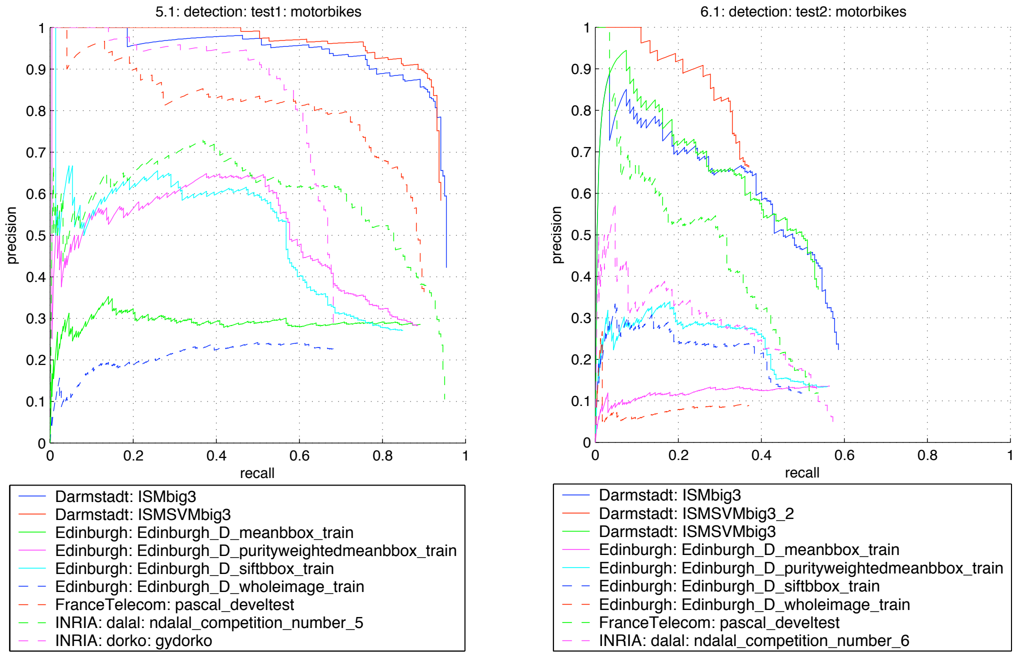


Figure 3.10: Results on the PASCAL Challenge motorbike detection task.

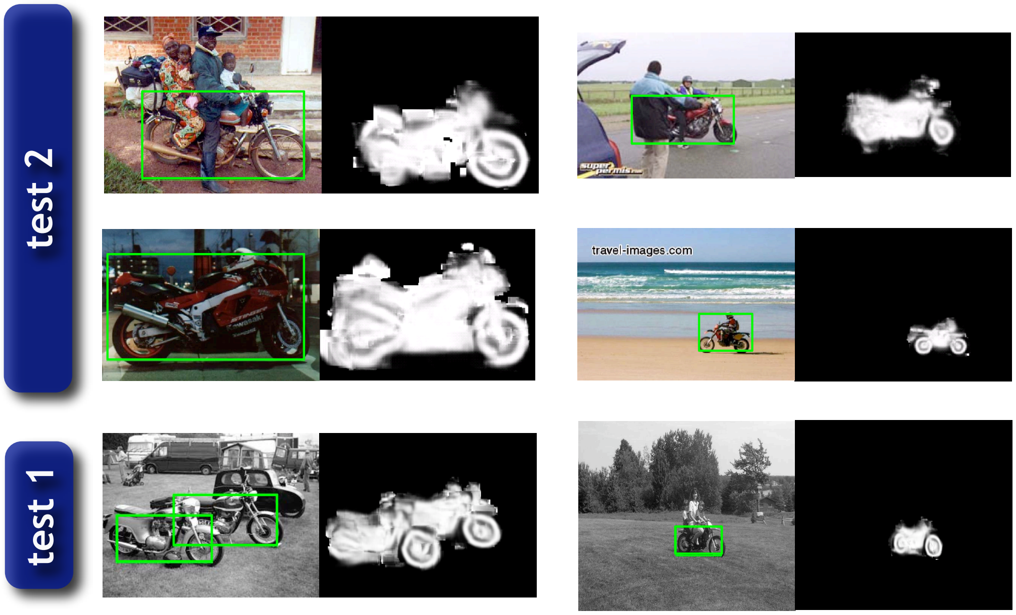


Figure 3.11: Example detections and segmentations on PASCAL Challenge motorbike detection task.

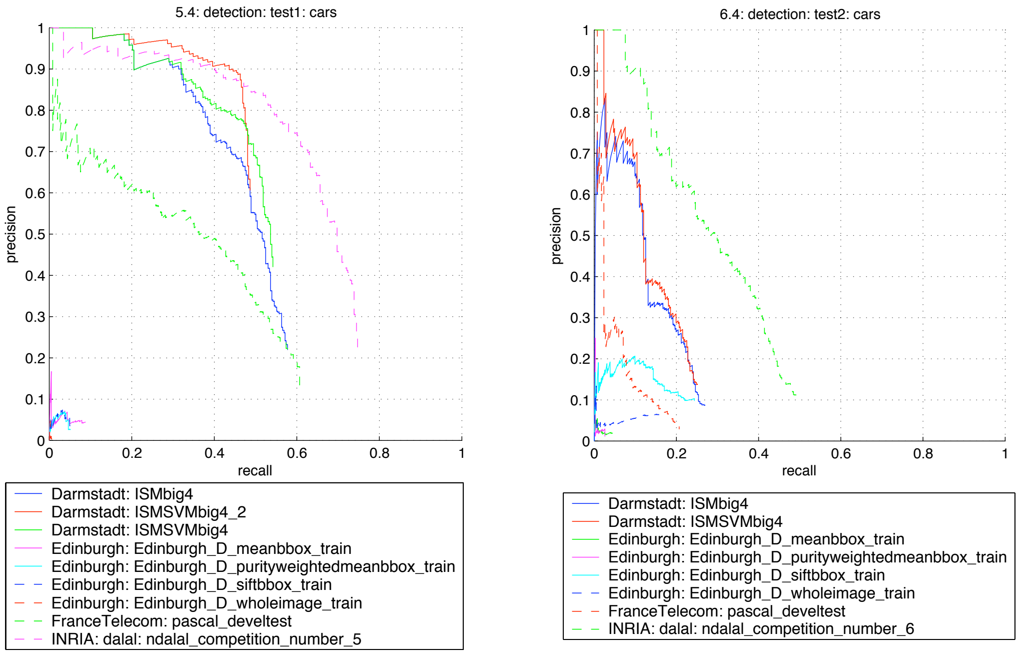


Figure 3.12: Example detections and segmentations on PASCAL Challenge car detection task.

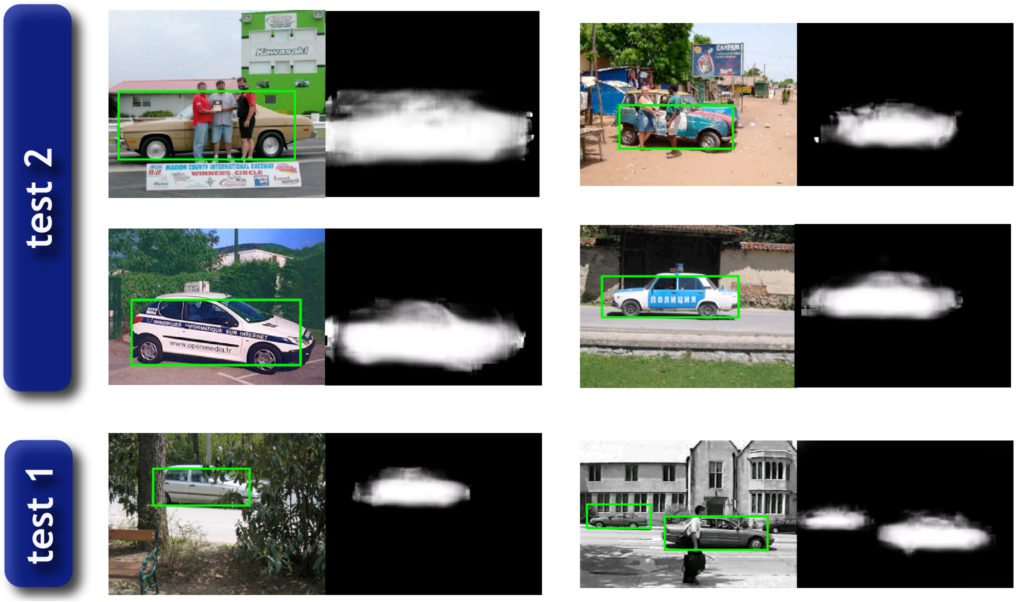


Figure 3.13: Results on the PASCAL Challenge car detection task.

Discussion. As can be seen from the comparison to our competitors in the PASCAL VOC 2005 detection challenge in Figure 3.10 and Figure 3.12, the ISM as well as our IRD approach outperformed the other approach on both motorbike test sets. On the car sets, we were second after the approach Dalal and Triggs (2005). We account the superior performance of the HOG approach on the car data to the capability of the histogram grid based approach to fine tune to rigid structure of the cars. The IRD consistently outperforms the ISM approach.

In addition to the detection competition, we also submitted results to the presence/absence competition. In contrast to the detection task, no localization of the object of interest is required. Basically, the task reduces to a classification problem if at least one object instance of the category of interest is present or not. While our approach worked well on the detection tasks, the results we derived based on the detections for the presence absence task were mediocre. This can be interpreted in two ways. First, although the detection approach can tell, why a certain image gets labeled with “car present”, it solves a more complicated intermediate problem. The introduced overhead and more difficult estimation problem can cause worse results, than directly aiming for the simpler target function. Second, it is not obvious how and if the excellent performance of algorithm designed for the presence/absence task translates to the detection problem. Depending on the approach, figuring out on which part of the image the actual decision is based on is not straight forward. Additionally, these approaches don’t necessarily base their decision on the object itself. E.g. detecting a road (given a particular data set) can already be considered strong evidence, that a car might be present in the image.

3.4 Conclusion

Summarizing our approach, we integrated a representative object detection method with a discriminative verification stage for the generated hypotheses. Both stages operate on a common codebook representation. They share and reuse the same information from sampled image patches, but interpret it in different ways. The ISM hypothesis generation stage searches for agglomerations of image patches that are locally consistent with a common object center. Treating each sampled patch independently, it can interpolate between different training examples and adapt to novel objects and changed articulations. The SVM verification stage, on the other hand, enforces stronger spatial constraints and verifies the global feature configuration. At the same time, its discriminative capabilities obviate the need for a dedicated background model, which is difficult to estimate reliably in a probabilistic framework. Finally, the tight integration with the output of the ISM stage removes the influences of translation and scale changes, which greatly simplifies the discrimination problem.

Together, both stages manage to reduce the number of false positives and cross-category confusions significantly and perform considerably better than either stage alone. In our experiments, we have shown this improvement both for a four-class detection/discrimination task and for object detection on two challenging data sets containing large scale changes and partial occlusion.

In particular, our submission to the PASCAL 2005 VOC challenge outperformed its competitors in the motorbike detection task and performed second best for the car data sets.

4

Weakly Supervised Learning via Scale-Invariant Patterns

The previous chapter described a method for object category detection that combines high recall with improved accuracy. However, this performance levels comes at the cost of high annotation effort. In particular, the ISM detector uses pixel-level segmentation of the training instances to leverage from the interleaved segmentation and detection framework. This is in accordance with the more general observation that high performance object detection methods tend to be trained in a supervised manner from relatively clean data. In order to deal with a large number of object classes and large amounts of training data, there is a clear desire to use as little supervision as possible.

This chapter proposes a new approach for weakly supervised learning of visual categories based on a scheme to detect reoccurring structure in sets of images. The approach finds the locations as well as the scales of such reoccurring structures in a weakly supervised manner. In the experiments those reoccurring structures correspond to object categories which can be used to directly learn object category models. Experimental results show the effectiveness of the new approach and compare the performance to previous fully-supervised methods.

The central problem addressed in this chapter is to discover and learn objects category models as reoccurring patterns of local appearance in sets of training data. It may seem quite unrealistic to discover object categories in this way. However, many appearance-based approaches explicitly or implicitly rely on the fact that both the local appearance as well as its structural layout exhibit reoccurring patterns that can be learned and modeled (e.g. Weber *et al.* (2000); Fergus *et al.* (2003); Leibe *et al.* (2004); Felzenszwalb and Huttenlocher (2005)). A key idea of our approach is therefore to discover reoccurring patterns in multiple images without the model of any particular object. Finding the locations and scales of such reoccurring structures effectively corresponds to unsupervised annotations of the training data. As we will show, the proposed approach enables effective object class discovery in unlabeled images. Using those estimated annotations a model of an object class can be learned.

The chapter is organized as follows: Section 4.1 describes a method for locating reoccurring structure for which in Section 4.2 we present a method to robustly

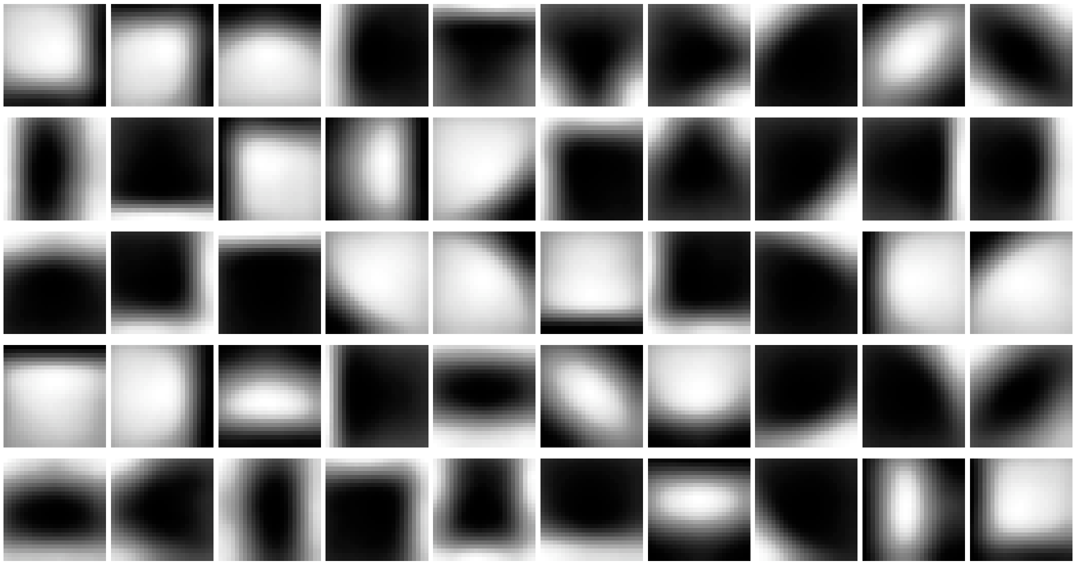


Figure 4.1: Cluster centers of the generic codebook. The codebook is unspecific since no knowledge about the particular object class is assumed during object discovery. 50 codebook entries are obtained by k-means clustering on the Caltech background dataset using a Hessian-Laplace interest point detector and gray value patches as descriptors.

estimate the intrinsic scale of the associated objects. Section 4.3 shows how a model like Leibe *et al.* (2008) can be learnt from the estimated annotations. Finally, we show in Section 4.4 the usefulness of the obtained information on a image ranking and an object detection task.

4.1 Object Discovery

Our new approach to unsupervised object discovery is based on efficiently finding reoccurring spatial patterns of local appearances. We use a generic codebook representation which is presented in Section 4.1.1. This representation is then employed to find locations (Section 4.1.3) and scales (Section 4.2) of reoccurring patterns in a set of training images.

4.1.1 Generic Codebook Representation

As in the previous chapter, we use an initial clustering procedure to obtain a visual codebook. Since we do not want to assume a priori information on parts or common structure of the object category, we use a generic codebook produced on unrelated background images. We extract image patches on the Caltech background images (as also used in Fergus *et al.* (2003)) using the scale-invariant Hessian-Laplace interest

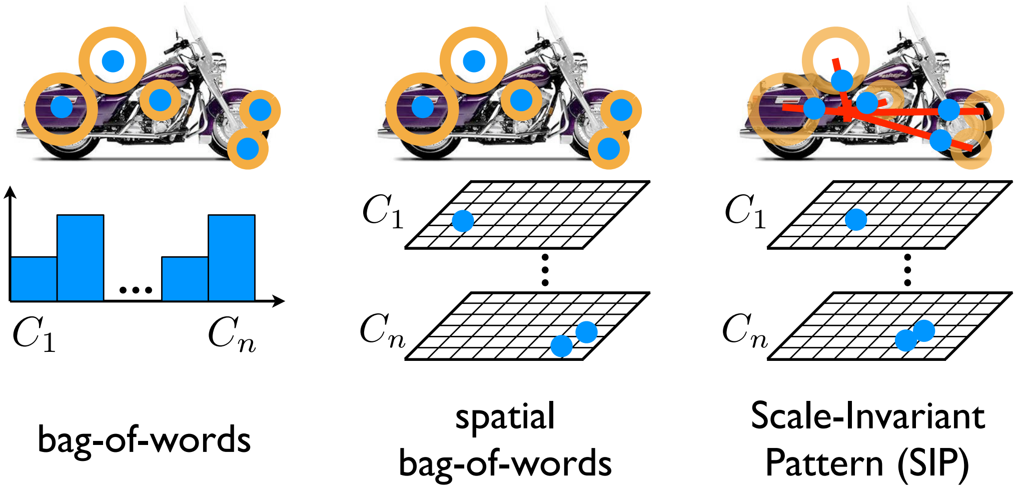


Figure 4.2: The scale-invariant representation we use is derived from the common bag-of-words representation. Recently a spatial resolution was added to each bin. We improve the robustness of these type of representations by adding a scale-normalization step yet retaining the spatial information which yields what we call a Scale-Invariant Pattern.

point detector of Mikolajczyk and Schmid (2005). Those image patches are clustered by k-means using normalized gray-scale correlation as similarity measure. Figure 4.1 shows the generic codebook with 50 cluster centers that we use for the object discovery.

4.1.2 Scale-Invariant Patterns (SIPs)

Given a set of K images with R_k reference points r in each image k, we define a pattern $\Psi_{k,r}$ in image k at reference point r to be characterized by a set of distributions $\{p(h|\Psi_{k,r,c})|c = 1, \ldots, C\}$. Each of the $p(h|\Psi_{k,r,c})$ encodes the spatial distribution of the features in image k that match to a certain codebook c, assuming a total codebook size of C. The coordinates $h = (h_x, h_y)$ are scale-normalized with the intrinsic feature scales σ (obtained from the scale-invariant interest point detector) and computed relative to a reference point $r = (r_x, r_y)$

$$h = \left(\frac{x - r_x}{\sigma}, \frac{y - r_y}{\sigma} \right). \tag{4.1}$$

As visualized in Figure 4.2, this representation can be seen as a extension to the bag-of-words representation of Csurka *et al.* (2004) by adding two spatial dimensions to each bin like in Lazebnik *et al.* (2006). But in contrast to previous approaches, we are re-gaining scale-invariance by normalizing the local feature configuration with respect to the center point.

Using this scale-normalized coordinates is beneficial, as the patterns become scale-invariant and characteristic for the particular reference point r. This allows to locate reoccurring patterns even though they appear at different global scales in the training images.

4.1.3 Method

We formulate the unsupervised discovery of reoccurring spatial patterns of local appearances as finding for each image the most likely pattern given all observed patterns in the training data. Therefore we are interested in finding the reference point $\hat{q}_j$ associated with the most likely pattern in each image j given all observed patterns $\mathbf{\Psi} = \{\Psi_{k,r} | k = 1, \ldots, K; r = 1, \ldots, R_k\}$

$$\hat{q}_j = \arg\max_q p(\Psi_{j,q} | \mathbf{\Psi}). \tag{4.2}$$

To simplify notation, the reference points q and r are assumed to be quantized. In fact, in the experiments the reference points will be defined on a regular grid overlaying the images. The likelihood estimate is obtained by marginalizing over the codebook entries c, scale-normalized coordinates h, reference points r, and images k

$$p(\Psi_{j,q} | \mathbf{\Psi}) = \sum_c \sum_h \sum_r \sum_k p(\Psi_{j,q,c} | h) p(h | \Psi_{k,r,c}) p(\Psi_{k,r,c}).$$

Using Bayes' formula we obtain

$$p(\Psi_{j,q,c} | h) = \frac{p(h | \Psi_{j,q,c}) p(\Psi_{j,q,c})}{p(h)}. \tag{4.3}$$

By assuming uniform priors, $p(\Psi_{k,r,c})$ and $p(h)$ can be written as constant $\frac{1}{Z}$. This assumption is justified, by a uniform partitioning of our data using k-means clustering. Eq. 4.3 simplifies to

$$p(\Psi_{j,q} | \mathbf{\Psi}) = \frac{1}{Z} \sum_c \sum_h \sum_r \sum_k p(h | \Psi_{j,q,c}) p(h | \Psi_{k,r,c}). \tag{4.4}$$

An example of this likelihood estimate on the multi-scale TUD motorbikes (available in the PASCAL database collection Everingham *et al.* (2005a)) is visualized as isolines on top of the images in Figure 4.3. The presented dense version of the likelihood, that is only defined point-wise on the reference points, was obtained by interpolation. This corresponds to a smoothing operation to increase the support of $p(\Psi_{j,q} | \mathbf{\Psi})$. In the visualization we can clearly see two maxima which correspond to two occurrences of the motorbikes.

Eq. 4.4 can be interpreted as collecting evidence for pattern $\Psi_{j,q}$ with respect to all other patterns $\mathbf{\Psi}$ by searching for matching feature with appearance c and scale-normalized position h. Although this seems computationally infeasible, we introduce an efficient method to evaluate eq. 4.4 using scale-invariant feature hashing - similar to the idea of geometric hashing of Wolfson and Rigoutsos (1997). The idea is to

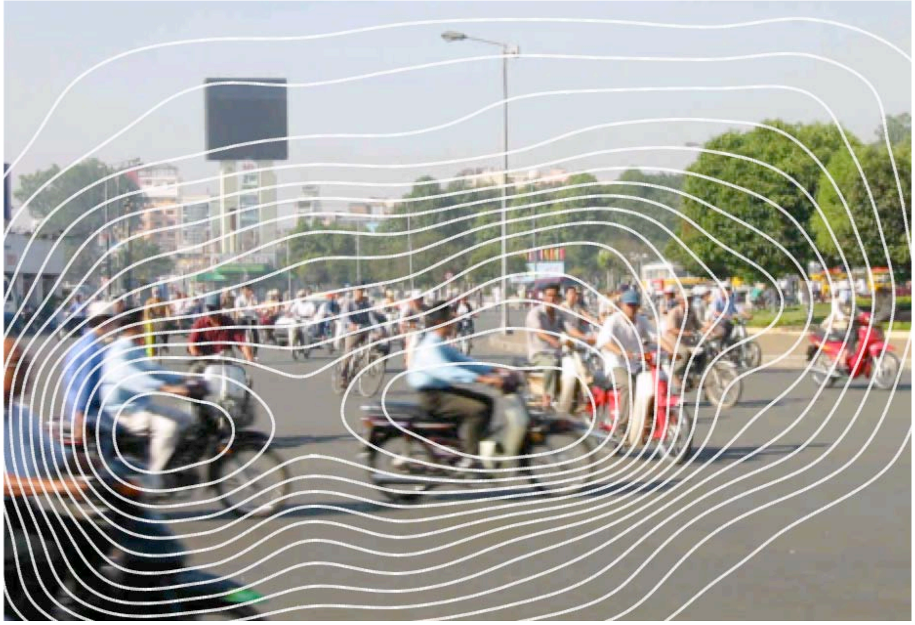


Figure 4.3: Example of the computed likelihood on the multi-scale TUD motorbikes.

index all features of the image database by quantized scale-normalized coordinates h and matching codebook clusters c and store them in the hash $\mathcal{H}$. Features which are matching to the same codebook and have similar scale-normalized position h are now stored in the same hash bin and considered to be matched. More importantly, the matches can be used to backproject the support of all patterns $\Psi_{j,q}$ with respect to all patterns. As a result, all $p(\Psi_{j,q}|\boldsymbol{\Psi})$ given by the complex eq. 4.4 can be computed by a single loop over the hash bins of hash $\mathcal{H}$.

4.1.4 Evaluation

To test the proposed procedure for object discovery with respect to robustness against translation, scaling, occlusion, and background clutter we ran tests on three object categories: motorbikes, cows, and cars. For the cows we use the training set of the TUD cows (again available form the PASCAL database collection) as well as the cows from Hillel *et al.* (2005). For the cars we use the training set of the PASCAL challenge (Everingham *et al.*, 2005a). Examples for the estimated object centers are shown in Figure 4.4. Despite the strong variations in appearance and view-point, the objects are successfully localized. The reference points r are quantized on a 10×10 grid.

To gain more insights, we perform a quantitative analysis on the Caltech motorbike training set as used in Fergus *et al.* (2003) which consists of 400 images. We compute the pixel distance between our estimate and the center of the groundtruth bounding box annotation as well as the distance normalized by the object width. As the results in Figure 4.5 show, nearly all errors are below a normalized distance

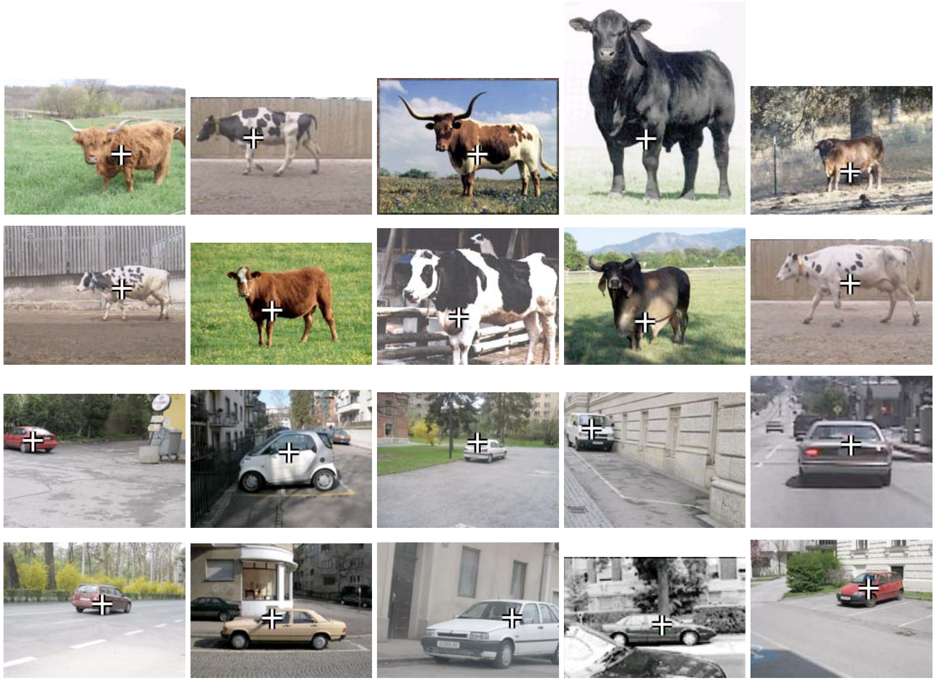


Figure 4.4: Example result of our procedure for object discovery on car and cow images including varying position, scale and viewpoint and heterogeneous background.

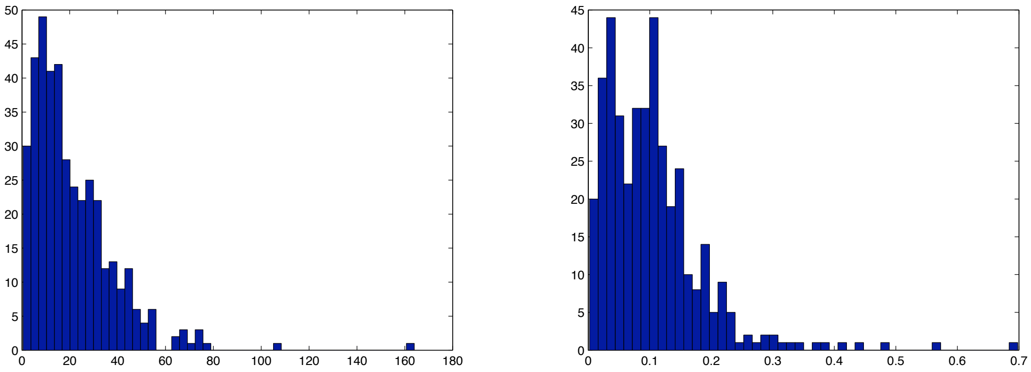
of 0.3, which corresponds to the first noise level we investigated in the analysis for the object scale estimation method in Section 4.2.2, that relies on this information. The average distance is much lower, namely 0.10 (again normalized distance). Since the groundtruth annotations are not really accurate themselves, the obtained error is considered to be low.

4.2 Object Scale Estimation

From the procedure for object discovery described in Section4.1 we obtain localized patterns $\Psi_{j,q}$ at reference points $\hat{q}_j$ for each image j. However, since these reoccurring patterns are obtained in a scale-invariant fashion, they are of unknown scale s. While it is advantageous, that no explicit knowledge of the object scale is required for discovering reoccurring patterns, tasks like training an object model for detection need an estimate of the object scale to learn a model across the training instances.

4.2.1 Method

The proposed method matches scale-invariant patterns to collect evidence for their associated global object scale. Different methods to obtain a robust estimate are proposed and evaluated. As the absolute, global object scale only exists with respect to a reference scale, we formulate the scale estimation problem as finding the pairwise



(a) (b)

Figure 4.5: Quantitative evaluation of the error in the center point estimation on the Caltech motorbike training set. (a) shows the distribution of the pixel distance between annotation and estimate and (b) shows the distribution of the distance normalized by the object width.

relative scale $\hat{\rho}_{k,l} = s_k/s_l$ between two discovered patterns Ψ_k and Ψ_l in a pair of images k and l. In analogy to eq. 4.2 we describe the problem of finding the most likely relative scale $\hat{\rho}_{k,l}$ with respect to the two patterns of the image pair as

$$\hat{\rho}_{k,l} = \arg\max_{\rho_{k,l}} p(\rho_{k,l}|\Psi_k, \Psi_l) \tag{4.5}$$

We assume that for matching features the ratio of the intrinsic scale σ of the matched structures is equal to the ratio of the global scales s between the patterns and their associated objects $\rho_{k,l} = s_k/s_l = \sigma_k/\sigma_l$. Accordingly, we expand eq. 4.5 and marginalize over the codebook entries c and the scale-normalized coordinates h

$$\begin{aligned}
p(\rho_{k,l}|\Psi_k, \Psi_l) &= \sum_{\sigma_l} p(\sigma_k|\Psi_k)p(\sigma_l|\Psi_l), \text{ where } \rho_{k,l} = \sigma_k/\sigma_l && (4.6)\\
&= \sum_{\sigma_l} p((\rho_{k,l}\sigma_l)|\Psi_k)p(\sigma_l|\Psi_l) && (4.7)\\
&= \sum_{c}\sum_{h}\sum_{\sigma_l} p((\rho_{k,l}\sigma_l), h|\Psi_{k,c})p(\sigma_l, h|\Psi_{l,c})
\end{aligned}$$

We use the same hashing technique as described in Section 4.1.3 to efficiently retrieve matching features and their associated scale ratio. Our efficient data structure allows to compute all these likelihoods in one loop over all hash bins.

As visualized in Figure 4.6, the estimates from eq. 4.5 can be interpreted as a fully connected graph, where the patterns in the images are the nodes with associated unknown object scale s and the relative scales of the patterns ρ are attached to the edges. To make our method robust with respect to outliers, we compute confidence scores for all estimated relative scales. These are computed by indentifying

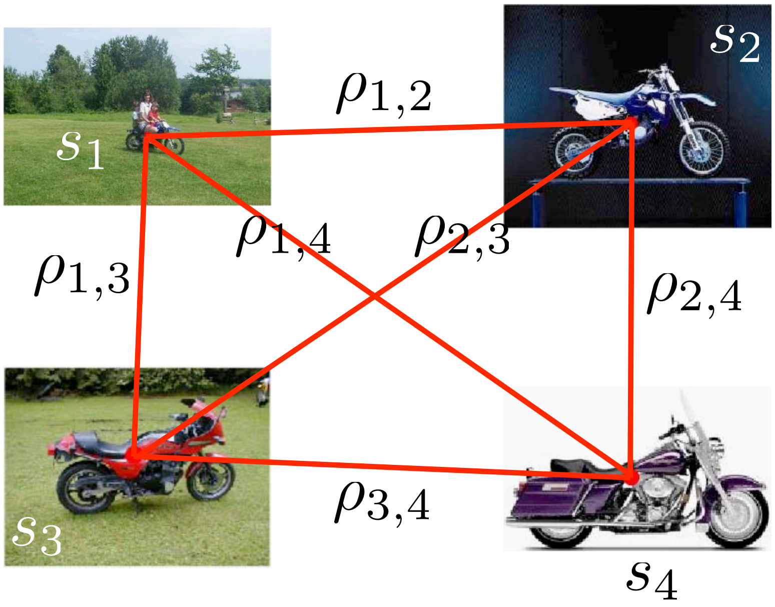


Figure 4.6: The dependencies between the unknown object scales s and the estimated relative scales ρ is visualized as a fully connected graph. After pruning of the graph, estimation of the object scales s are estimated by least squares.

image triplets with consistent relative scale estimates: Given three images I_a, I_b, I_c with their relative scales $\rho_{a,b}, \rho_{b,c}, \rho_{a,c}$, the confidence for all three scale estimates is increased if the equation $\rho_{a,b}\rho_{b,c} = \rho_{a,c}$ is fulfilled.

In this paper we investigate three different methods to derive a unique scale estimate for each pattern from the pairwise relative scale information: *least squares*, *maximum spanning tree*, and *min-linkage method*.

The *least squares method* is based on a linear system of equations to estimate the unknown scales without using the computed confidences. Considering two patterns Ψ_k, Ψ_l with the global scale of the patterns s_k, s_l of the associated object instances, we compute a least-squares fit for the global scales s from all the estimated relative scale according to:

$$\frac{s_k}{s_l} = \rho_{k,l} \implies \log s_k - \log s_l = \log \rho_{k,l}. \tag{4.8}$$

This leads to the linear system of equations for all N images:

$$\begin{pmatrix} 1 & -1 & 0 & & & \dots \\ 1 & 0 & -1 & 0 & & \dots \\ & & & \vdots & & \\ 0 & 1 & -1 & 0 & & \dots \\ 0 & 1 & 0 & -1 & 0 & \dots \\ & & & \vdots & & \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \log \begin{pmatrix} s_1 \\ s_2 \\ \\ \vdots \\ \\ s_{N-1} \\ s_N \end{pmatrix} = \log \begin{pmatrix} \rho_{1,2} \\ \rho_{1,3} \\ \vdots \\ \rho_{2,3} \\ \rho_{2,4} \\ \vdots \\ \rho_{N,N+1} \end{pmatrix} \tag{4.9}$$

This method is computational expensive, because the number of equations grows quadratically in the number of images, and its estimates are sensitive to outliers.

The *maximum spanning tree method* computes a maximum spanning tree on the graph of confidences. The scale estimates can be directly computed from this tree by fixing one scale. Although this method has low computational complexity, the estimates are rather unstable. This is confirmed by our evaluation in Section 4.2.2.

As a compromise between efficient computation and robust estimation, we propose a third method. The *min-linkage method* considers for every image the n most confident relative scales to all other images and therefore the number of equations grows only linearly with the number of images. The estimate of the scales is still robust due to the least-squares estimation.

The above described methods estimate relative scales, however, for the detection experiments (Section 4.4.2) an absolute scale based on the extent of the object is required. One possibility is to specify a reference scale for one image. In the experimental evaluation we use the following heuristic. The absolute object radius is chosen to be twice the mean feature distance to the center *after* aligning all objects with the computed relative scale estimates.

4.2.2 Evaluation

To evaluate the accuracy of our new scale estimation scheme, we again use the Caltech motorbike database with annotated object centers, which has a scale variation of 2 octaves. Figure 4.7 shows the groundtruth object scales as well as the different estimates for all 400 motorbikes. The object instances were sorted along the horizontal axis according to the groundtruth scale, which was derived from the width of the bounding box annotations. For the minimum linkage method, we chose to preserve $10\% = 40$ of the most confident relative scales. The methods computes reliable estimates of the object scale even though the presented object instances were spread over 2 octaves.

As we estimate the center-point in the final system using the procedure of Section 4.1, we investigate the robustness of our scale estimate with respect to noise added to the coordinates of the center-point. We run these experiments on the 400

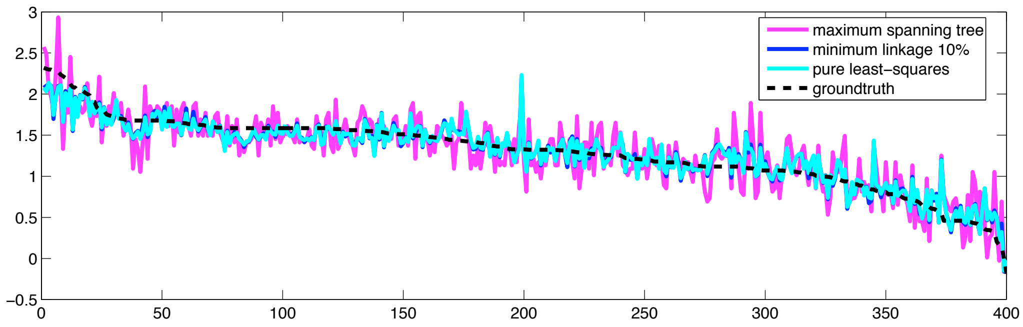


Figure 4.7: Evaluation of the scale estimate on the Caltech database. The 400 images are sorted by the true scale of the object. The scale estimate for each of the proposed methods is plotted logarithmically. The curve for the pure least-square and minimum linkage method are virtually identical.

noise level	least squares	max span tree	minLink 10%	minLink 1%
σ_0	0.0233	0.0579	0.0231	0.0271
σ_1	0.0425	0.1209	0.0431	0.0505
σ_2	0.0763	0.2489	0.0735	0.0975
σ_3	0.1447	0.4813	0.1375	0.1647

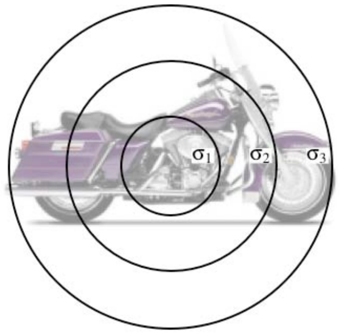


Table 4.1: Noise analysis of our new object scale estimation technique. Different methods to obtain the final estimate are evaluated at different noise levels of the annotation of the object center. The error in the scale estimate is given in mean squared deviation form the groundtruth in log-space.

Caltech motorbikes and add Gaussian noise with increasing variance to the center-point annotation. We chose three noise levels with standard deviations of $\frac{1}{6}$, $\frac{1}{9}$ and $\frac{1}{18}$ times the object width, whose 3σ-radius is visualized to the right of Table 4.1. σ_0 corresponds to the experiment without any noise added to the center-point annotation. We report the mean squared deviation in log-space from the correct scale, that is computed on the bounding-box width of the annotation provided by Fergus *et al.* (2003). As we are interested in the quality of the estimation of the relative scale, we remove the bias in the estimation.

The results in Table 4.1 show that the approaches using least-squares fitting outperform the maximum spanning tree solution. The graph reduction of the minimum linkage method gives only minor improvements at high noise level. Nevertheless we favor the minimum linkage method, due to its much lower computational complexity.

Even when the noise is amplified until the 3σ-radius reaches $\frac{2}{3}$ of the object radius - which is twice the noise level we measured for the center point estimate in Section

4.1.4 - the mean deviation of the estimated scales from the true scale is roughly $\frac{1}{4}$ of an octave for least-squares and minimum linkage. In particular, the method is robust enough to handle the noise in the center-point annotation produced by the estimation proposed in Section 4.1, as it is below the first noise level as shown in Section 4.2.2.

As a conclusion we use the minimum linkage method in our following experiments, as it shows about the same accuracy as the full least-squares, but with a much lower computational cost.

4.3 Model Estimation for Detection

As described in Section 2.2.1, the *Implicit Shape Model (ISM)* (Leibe *et al.*, 2008) is a versatile framework for scale-invariant detection of object categories, which has shown good performance on challenging detections tasks like the PASCAL challenge. It uses a flexible non-parametric representation for modeling visual object categories by spatial feature occurrence distributions with respect to a visual codebook. Additionally the method allows for back-projecting the support of the hypotheses to infer figure-ground segmentation masks and performing an MDL-based reasoning to resolve multiple and ambiguous hypotheses (Leibe *et al.*, 2004). However, the generation of an object specific visual codebook and the MDL-based reasoning step require figure-ground segmentations for the training images which introduce high annotation effort. One of our contributions is to show that one can achieve high recognition performance by using the estimated center point (Section 4.1) and scale (Section 4.2) instead of manually produced segmentations. As we do not have a detailed segmentation mask at our disposal when using those location and scale estimates, we use a much simple but (as will be seen in the experiments) effective approximation. Figure 4.9 shows our rough approximation by assuming the segmentation to be a disc specified by the estimated center point and scale.

To learn the ISM model, we first switch from the generic codebook (Section 4.1.1) to an object specific representation using SIFT descriptors (Lowe, 2004) computed on Hessian-Laplace interest points (Mikolajczyk and Schmid, 2005). We use the approximated segmentation (discs) to determine the object features for clustering. Given the approximated segmentation and the new codebook, we can proceed training the ISM as described in Section 2.2.1. Despite the crude approximation of the segmentations with discs, it is possible to infer segmentations for the hypothesis on test images as shown in Figure 4.9.

4.4 Experiments

Whereas the previous sections analyzed the proposed object discovery and object scale estimation separately, this section shows the applicability of these components to two tasks - namely image ranking and object category detection. While the ranking task also show the scalability to large numbers of images, the detection experiments evaluate how the proposed method generalizes to different categories. In addition we will show that the approximated segmentation mask from Section 4.3 are effective and even crucial to obtain high level detection performance.

4.4.1 Image Ranking

In the following experiments, we show that the proposed method for unsupervised object discovery from Section 4.1 can be used on its own for an image ranking task. Using a keyword search for motorbikes we downloaded 5246 images containing a wide range of different motorbike types (e.g. cruiser, sport-bike, touring, scooter, moped, off-road, combination) captured from different viewpoints. Quite a number of those images only show close-ups, parts or even unrelated objects. Our task is to sort these images out. We use our method for object discovery to rank the images by the likelihood (eq. 4.4). Note, that this ranking is obtained in a totally unsupervised fashion and no validation set as in Fergus *et al.* (2005a) is needed. Figure 4.10 (left) shows the ROC curves obtained by running our approach with and without spatial information. If the spatial information of the features is discarded, out representation reduces to a bag-of-words representation. The use of spatial information improves the results significantly, which demonstrates the improvement of our model over purely appearance-based approaches. Qualitative results for our new approach using appearance and spatial structure are shown in Figure 4.8. As scooters were the dominating motorbike type in the set (1169 of 5246), they also appear first in the ranking.

4.4.2 Visual Category Detection Task

In the detection experiments we train a model according to Section 4.3 and use it to localize instances of an object category of interest in the test images. Detections are only accepted as correct if the hypothesized bounding box fits the groundtruth annotation. Also multiple detections are counted as false positives. For better comparability we use the acceptance criterion described in Section 3.3.3. We want to emphasize, that no parameters had to be tuned for the proposed approach for unsupervised learning. In terms of efficiency, the approach for object discovery can estimate object locations in 200 images in 11 Minutes on a 3Ghz Pentium4, whereas the object scale estimation takes 6 Minutes.

Unsupervised Learning for Detection Figure 4.10(middle) shows results on the multi-scale TUD motorbike test set. The model is trained on the Caltech motorbikes. Note that this test set includes significant scale variations, partial occlusions and multiple instances per image. For comparison we also include the results from Section 3.3.4 on this dataset of 80% EER using accurate pixel-level segmentation and ISM (supervised training with MDL) and 81% adding an additional SVM-stage

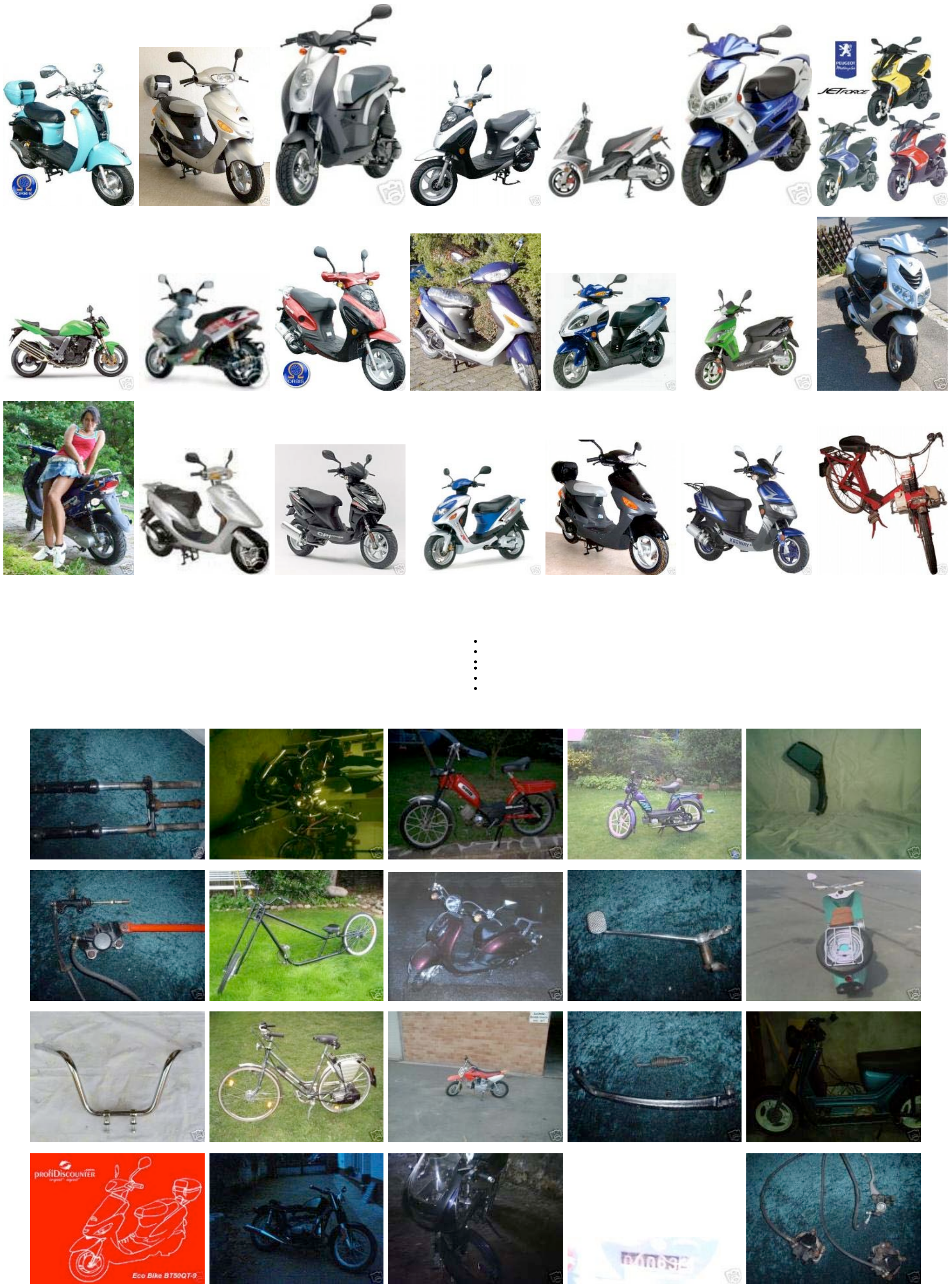


Figure 4.8: The proposed method for Object Discovery also facilitates ranking of the images. (top) best ranked images (bottom) worst ranked images.

(supervised training with MDL+SVM). Surprisingly, the performance of the proposed unsupervised object discovery method (specific SIFT codebook with MDL - #150) is very similar to the supervised training of ISM. The EER of 81% can be further increased to 84% by using 400 instead of 150 training images (again in an unsupervised fashion). Compared to the SVM approach of Chapter 3 the precision

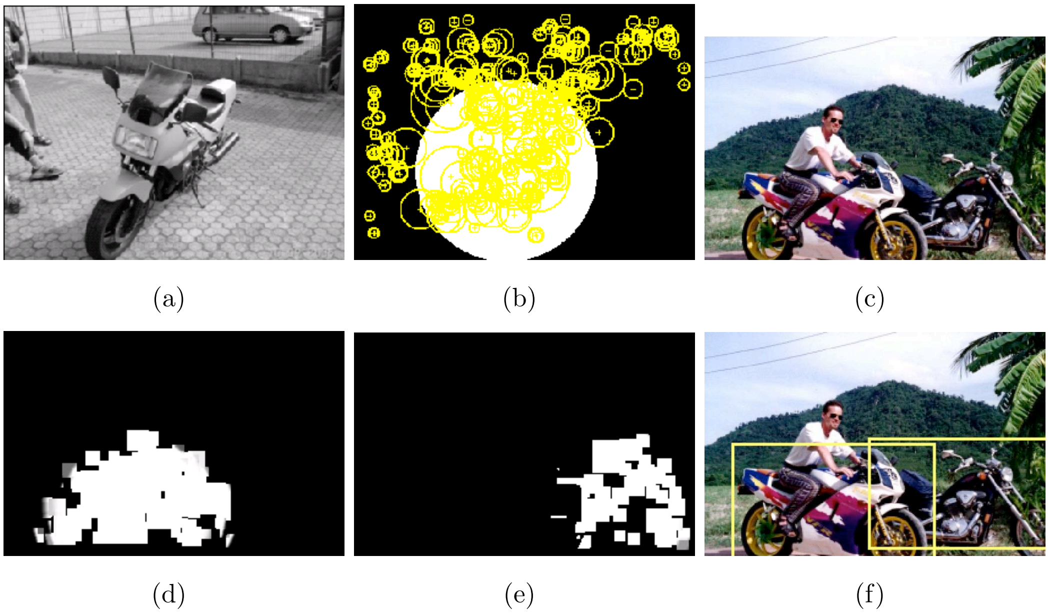


Figure 4.9: (a) training image (b) estimated approximation of object segmentation (c) test image (d)+(e) inferred segmentation for hypothesis (f) final detections

is slightly worse, but the achievable recall is higher. So adding an SVM classifier in a similar fashion has the potential to further increase the overall performance. Overall the results are highly encouraging as they indicate that high annotation effort can be replaced by using a larger amount of training data.

Evaluation of the Approximated Segmentation Masks Figure 4.9 shows a training image (a) with the approximate segmentation mask (b) obtained by Object Discovery and Scale Estimation, test image (c), inferred segmentation of hypothesis (d) and (e) for the final detections displayed in (f). While the approximated segmentation mask for the training example is far from being perfect, the inferred support of the hypotheses on the test image is approximately correct for both motorbikes.

Figure 4.10(middle) shows how the performance increases significantly when this approximation is used to perform the MDL-based hypothesis verification. The results support our claim, that the estimated segmentation masks are accurate enough and facilitate the training of a model that gives competitive performance. The figure also shows that switching form a generic codebook to the a object class specific SIFT codebook as described in Section 4.3 results in a major improvement.

Generalization to other Categories To investigate how this approach generalizes to other categories and to compare our method to previous work, we conduct experiments on cows, faces, and cars. The results are reported in Figure 4.10(right). The training sets TUD cows and Caltech faces are selected, as they include significant amount of variation of the object position in the training data to underline the

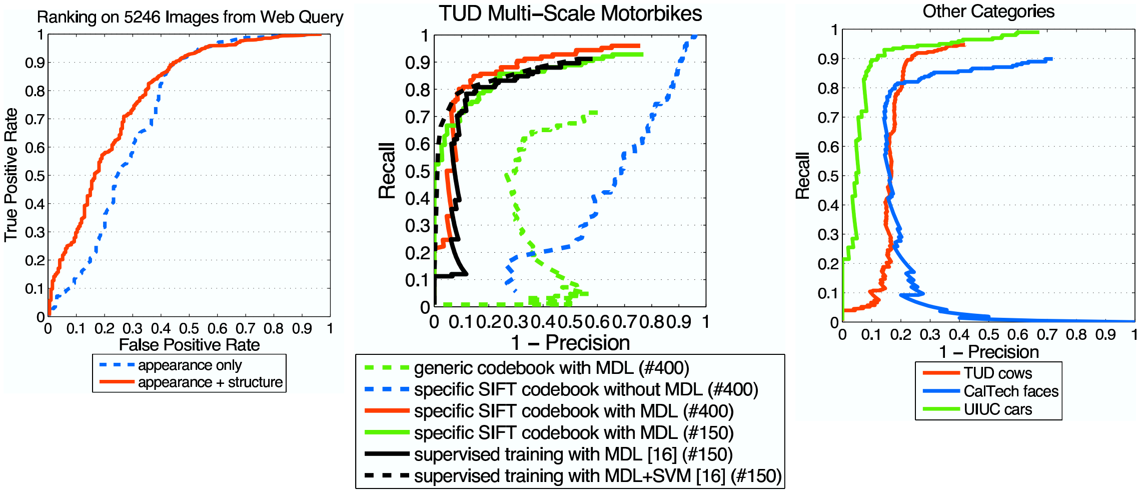


Figure 4.10: (left) ROC-curve of ranking task (middle) performance comparison to supervised baseline and performance increase due to approximated segmentations (right) generalization to other categories and data sets

performance of the proposed method for object discovery. For the cows we use the same test setting as in the supervised approach of Chapter 3. Our weakly supervised approach achieves an equal error rate performance of 79.9% whereas the supervised reference of Chapter 3 achieved 93.2% Fritz *et al.* (2005). As the background is for some training images the same, we learnt it as reoccurring structure. As it is part of the model, we get some strong hypotheses on the these background structures which also occur in the test set and that are responsible for the decrease in performance. On the UIUC car and caltech face database we also compare to the weakly supervised method of Fergus *et al.* (2003) On the cars we get an equal error rate performance of 89.5% in comparison to 88.5% in Fergus *et al.* (2003) using the same evaluation criterion. We achieve this performance training on only 50 car images and their mirrored versions from the TUD car database. The best performance on this dataset is reported by the supervised method in Leibe *et al.* (2004) achieving 97% equal error rate performance. In Fergus *et al.* (2005b) a detection performance of 78% equal error rate is presented on the caltech face database for the model of Fergus *et al.* (2003). Our approach achieves a significant improvement by an equal error rate performance of 81.1%.

4.5 Conclusion

We have proposed an efficient and flexible framework for discovering visual object categories in an weakly supervised manner which makes use of appearance and spatial structure at the same time. The approach is based on two new components for object discovery and object scale estimation, that extract information about reoccurring spatial scale-invariant patterns of local appearance. The experimental results show that our system facilitates weakly supervised training of an model for object class detection that has equal or even better performance than previous weakly supervised approaches. In addition, the method was used to rank images without any supervision or validation set. Results are presented on large image database of over 5000 images including a significant amount of noise. Finally, we obtained comparable results w.r.t. a strongly supervised detection system on a challenging multi-scale test set. We showed that we can compensate for the decrease in performance by adding more training examples.

5

Cross-Modal Learning at Different Levels of Supervision

The majority of today's object categorization methods use either supervised or unsupervised training methods. While supervised methods tend to produce more accurate results, unsupervised methods are highly attractive due to their potential to use far more and unlabeled training data. On the other hand, purely unsupervised approaches have inherent limitations on what can be learnt only from data given a particular notion of visual similarity. This chapter proposes a novel method that uses unsupervised training to obtain visual groupings of objects and a cross-modal learning scheme to overcome those inherent limitations of purely unsupervised training. The method uses the scale-invariant object representation by scale-invariant patterns from Chapter 4 that allows to handle labeled as well as unlabeled information in a coherent way. One of the potential settings is to learn object category models from many unlabeled observations and a few dialogue interactions that can be ambiguous or even erroneous. We explore a tutor driven learning scenario and first experiments demonstrate the ability of the system to learn meaningful generalizations across objects already from a few dialogue interactions.

5.1 Tutor-driven Learning Scenario

As outlined in Section 2.1.3 methods from previous work use different learning paradigms ranging from supervised methods like Leibe *et al.* (2005), over weakly supervised methods like Fergus *et al.* (2003) to unsupervised methods like Sivic *et al.* (2005); Grauman and Darrell (2006). Following common practice, these systems are evaluated on predefined training and test sets enabling direct comparisons. However for interactive and tutor-driven learning scenarios, as investigated in the area of cognitive vision systems, it is highly important that models and representations are flexible and evolvable over time enabling continuous or even life-long learning. This goal is not only much harder to achieve but it is also more difficult to evaluate and compare. Consequently it is not clear how the above mentioned approaches could be extended to deal with this more challenging scenario.

As we understand cognitive vision systems, one of their most important and fundamental abilities is to evolve over time by actively and passively acquiring new knowledge and incorporating that knowledge into the system. While there exists a wide range of sources of knowledge, in this chapter we focus on the ability to acquire new knowledge through dialogue interactions with humans. In this scenario we can identify a number of requirements a cognitive vision system needs to fulfill. First, to enable interactivity, the representations and models of the systems need to enable incremental processing and learning. Secondly and in order to test and evaluate such systems all processing should be done in real-time or at least at speeds that allow real interactivity. Third, as humans will use language to interact with the system, the learning mechanisms have to allow cross-modal learning from vision and language. And forth the learning algorithms should enable to deal with ambiguous and even erroneous input both from vision and language.

In the following, we present an approach for cross-modal learning of visual categories which integrates language and vision input. Language provides "scene descriptions", describing objects and their spatial relations in a given scene, which provides a top-down description which is then related to the bottom-up generalizations of the vision system. The scene descriptions can be interpreted on ontologically rich knowledge representations, which make it possible to use ontologies to mediate between linguistically expressed meaning and the categories formed in the visual system. Using the hierarchical structure of ontologies, and the possibility to perform ontological inference over instances on these ontologies, provides a more general and better scalable approach to "visual grounding" of language than provided by the string-based approach proposed in Roy (2002), or earlier ontology-based approaches such as Kruijff *et al.* (2006a).

5.1.1 Related Work

How entities in the real-world can be related to abstract concepts is subject of study in diverse fields of research such as psychology, computational linguistics and artificial intelligence. Commonly, this process is referred to as the symbol grounding problem Harnad (1990). Our approaches is related to the work of Roy (2002) and Steels (2008) who presented approaches to associate spoken utterances with visual percepts. In contrast, we go beyond single utterances by combining dialogue and vision. This makes it possible to incrementally provide information to learning visual categories etc, rather than"requiring" that all info comes in a single utterance. In addition we use a hierarchical structure of ontologies with the possibility to perform ontological inference over instances on these ontologies. We argue that this provides a more general and better scalable approach to "visual grounding" of language than provided by the string-based approach, or earlier ontology-based approaches such as (Kruijff *et al.*, 2006a).

Socher *et al.* (2000), Bauckhage *et al.* (2001) and Skočaj *et al.* (2007) implement interactive systems with a spatial reasoning scheme similar to ours. While Socher *et al.* (2000) and Bauckhage *et al.* (2001) only model a fixed set of object instances, we model representations of visual categories that are created at run-time and evolve

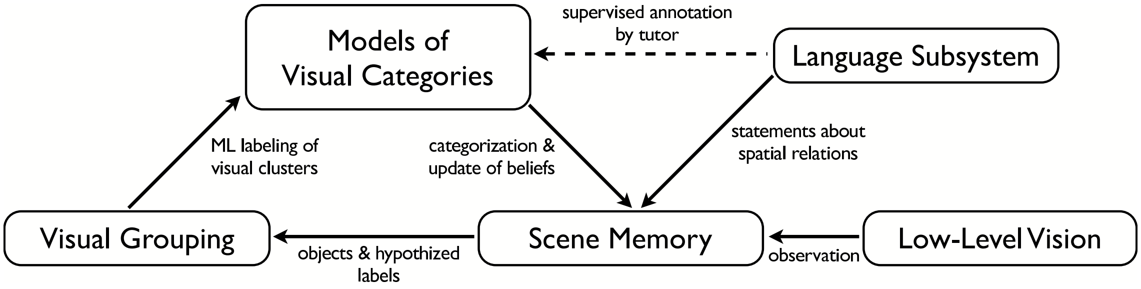

Models of Visual Categories
supervised annotation by tutor
Language Subsystem
ML labeling of visual clusters
categorization & update of beliefs
statements about spatial relations
Visual Grouping
objects & hypothized labels
Scene Memory
observation
Low-Level Vision

Figure 5.1: System overview.

over time. On the other hand Skočaj *et al.* (2007) does learn visual concepts, but not at the level of visual categories.

Steels and Kaplan (2001), Arsenio (2004) and Kirstein *et al.* (2005), target specifically interactive and incremental learning for a robotic platform, but non of them deals with category-level learning and recognition.

Previous ideas on reducing the number of required training examples for learning a visual category are largely based on priors over model parameters in a Bayesian Learning setting (e.g. Fei-Fei *et al.* (2003b)) or ideas from transfer learning (e.g. Bart and Ullman (2005)). However, these approaches do not fit our computational real-time constraints and their applicability to the object domain we are interested in (office/household) is questionable.

System overview. Figure 5.1 shows an overview of the presented system, which is tightly related to the structure of this chapter. Section 5.2 describes the vision system which is decomposed into low-level functionality (feature extraction, object discovery and object representation (Sec. 5.2.1)), the unsupervised visual grouping step (Sec. 5.2.3) and the categorization procedure (Sec. 5.2.4). Section 5.3 describes the language system that parses an utterance to a logical form. Section 5.4 explains the spatial reasoning processes that associate the expressions with the visual observation which are then used to probabilistically associate labels to the clusters obtained from visual grouping. Section 5.5 illustrates the functions of the integrated system and provides empirical evidence for our claims.

5.2 Vision Sub-System

The description of the vision sub-system is divided into three parts. First, the low-level functionalities are described, that extract local image features, discover object centers and extract sparse scale-invariant patterns at the hypothesized object positions. Second, the visual grouping procedure is explained, that provides the system with a data-driven generalization over category instances that is obtained in a totally unsupervised manner. This is implemented by an agglomerative clustering on the extracted scale-invariant patterns. Third, the model for performing categorization is described that is based on the same scale-invariant representation and can handle information obtained in a supervised, semi-supervised and unsupervised fashion.

5.2.1 Object Representation by Sparse Scale-Invariant Patterns

We build on the scale-invariant pattern (SIP) representation described in Section 4.1.2. This representation is a key ingredient in our system as it provides an embedding of visual patterns and objects into a vector space that facilitates clustering and recognition in a cheap yet effective manner. First, we describe the low-level feature description we base our representation on and then describe how the representation translates to the targeted application scenario.

When a new image is grabbed from the camera, SIFT descriptors (Lowe, 1999) are extracted at hessian-laplace interest points (Mikolajczyk and Schmid, 2003). While there exists a wide range of interest point and descriptor combinations, we opted for this particular combination based on evaluations on different categorization tasks in Mikolajczyk *et al.* (2005). Following a common philosophy in the field to visual categorization (Csurka *et al.*, 2004; Leibe *et al.*, 2005; Sivic *et al.*, 2005; Lazebnik *et al.*, 2006; Agarwal and Triggs, 2006; Fritz and Schiele, 2006), we first generate a visual codebook based on clustering of detected features. We use a codebook with 1000 entries obtained by k-means clustering. The matching procedure is accelerated by re-normalizing the codebook entries c (corresponding to the cluster centers) to a fixed length, which transforms the L_2-distance to a feature $\vec{x}$ into a scalar product with an additive constant:

$$||\vec{x}-\vec{c}||^2_{L_2} = (\vec{x}-\vec{c})^2 = \underbrace{\vec{x}^2+\vec{c}^2}_{const} - 2\vec{x}\vec{c} \tag{5.1}$$

The matching of all image features to all codebook entries can now be computed by a simple matrix multiplication which lends itself to further speed-ups. The introduced errors are found to be negligible.

Based on this feature representation we compute the scale-invariant patterns Ψ as described in Section 4.1.2. This representation is the basis not only for discovering objects in the scene, but also for visual grouping and object categorization. For efficiency, we store these patterns as sparse vectors Ψ. Sparsity is one key property that leads to the efficiency we are aiming for while still maintaining a descriptive representation of the visual input.

To provide a first intuition, that the embedding implied by this representation can be effective for our task and distances in that space make sense in terms of visual similarity, Figure 5.2 shows a multi-dimensional scaling plot of samples from 5 categories (mobile, pen, bottle, can, apple) in that space. The plot shows that 4 of the 5 classes are already well separated without using any label information. The confusion between cans and bottles is rather reasonable - in particular given the small statistic we provided. On the right, we present an example of an embedding of 4 categories from the Caltech database - namely airplane, face, leave,

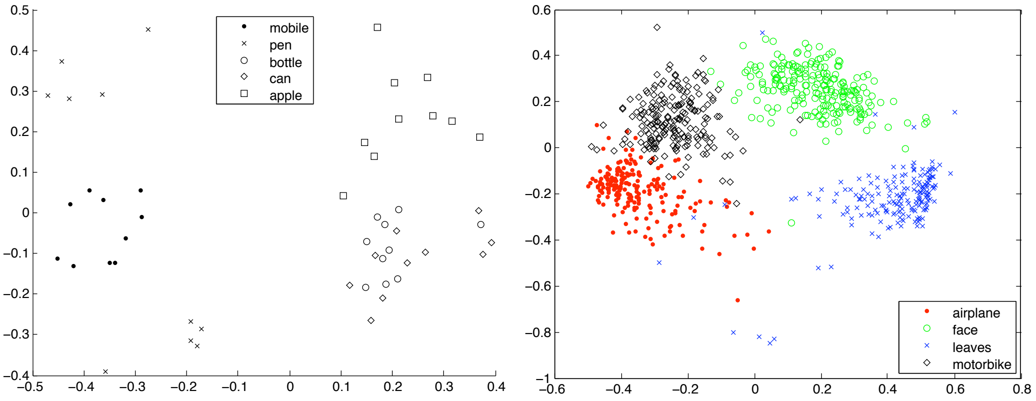


Figure 5.2: Multi-dimensional scaling using scale-invariant pattern representation. (left) 4 of the 5 classes (mobile, pen, bottle, can, apple) are clearly separated without using any supervision and despite the small statistic. Only the categories bottle and can are confused. (right) all 4 caltech classes (airplane, face, leave, motorbike) are well separated using 200 examples from each category. The groundtruth labels are displayed color-coded.

motorbike. These images are a subset of the images used in Sivic *et al.* (2005) supplemented by the leaves from the Caltech database. Again we use the same strategy by encoding the category instances as scale-invariant patterns and applying multi-dimensional scaling. Given approximately 200 samples from each category, we have a greatly improved statistic that results in a surprisingly good separation given the low-dimensional embedding, variation in the data and the unsupervised approach. To quantify our findings, we cluster the instances in the original space by k-means clustering. Simply choosing 4 cluster centers and assigning each cluster to the category that is most prominent in that cluster yields already an average accuracy of 94.61%. As this approach is dependent on the initialization of the k-means clustering, we re-run the experiments 5 times, but didn't observe any significant change in the obtained accuracy.

5.2.2 Object Discovery

We adopt the technique of discovering objects in the scene as reoccurring patterns described in Section 4.1. The method can be seen as an adaptive approach for acquiring a feature statistics of objects in recently observed scenes. This statistic is used to hypothesize object centers q_j in scene j given uniformly sampled scale-invariant patterns $\mathbf{\Psi}$ from previous scenes by finding maxima of the computed likelihood function:

$$\hat{q}_j = \arg\max_q p(\Psi_{j,q}|\mathbf{\Psi}), \tag{5.2}$$

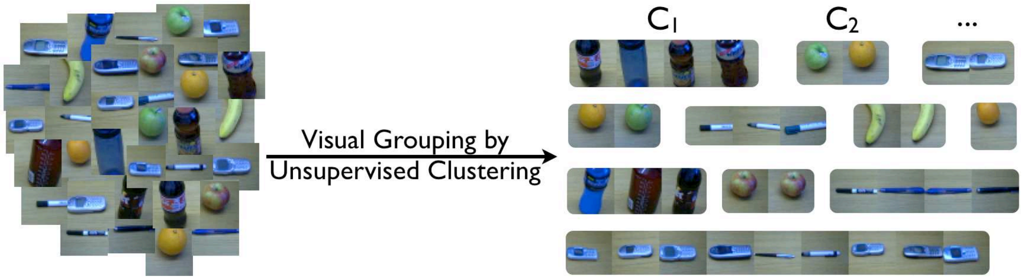


Figure 5.3: Visual grouping of objects by clustering.

where the query patterns $\Psi_{j,q}$ are again uniformly sampled in the image of interest. Still this would mean marginalizing out all the spatial and codebook bins in the scale-invariant patterns, which would lead to high computational costs. But due to the sparsity of our representation we can make use of efficient hashing related to the approach applied in geometric hashing (Wolfson and Rigoutsos, 1997). As detailed in Section 4.1.3, the similarity score to all scale-invariant patterns can be obtained by a single sweep over the non-zero bins of the sparse query pattern $\Psi_{j,q}$. Instead of taking only the global maxima of this likelihood function, we now hypothesize all local maxima as object centers in order to deal with multiple instances presented in one scene.

As a result, an observed scene is represented by a set of scale-invariant patterns Ψ at the locations indicated by this discovery procedure. By keeping a fixed number of patterns (the most recent ones) in memory, this online method for object discovery meets our real-time requirements and also consumes a constant amount of memory. Even though the original method can handle significant background clutter we cannot benefit from that large statistics in our target scenarios, as the number of interactions with the system is limited. In a sense we decided to trade generality for real-time capability of the system. As it has been shown that arbitrary backgrounds can be handled when sufficient statistics are available, one could extend our system in this direction.

5.2.3 Unsupervised Visual Grouping

Similarly to Grauman and Darrell (2006), we use an agglomerative clustering scheme (average linkage) to group object instances in an unsupervised manner. The objects are represented as scale-invariant patterns (Sec. 5.2.1) which we will denote as sparse vectors Ψ in the following formulas. As described before we normalize to unit length and we use the scalar-product to measure similarity between these objects. We prefer to use agglomerative clustering over k-means as we do not want to specify the number of visual clusters a priori. The threshold required for the agglomerative clustering scheme is set empirically to a constant value for all our experiments. Figure 5.3 visualizes the clusters C_1 to C_N obtained by our system given the observed objects displayed on the left. Although there are some confusions, we observe a good generalization across category instances. In order to obtain representatives $\bar{\Psi}_{C_l}$ for each cluster C_l, we compute a weighted sum of the observed patterns Ψ_k:

$$\bar{\Psi}_{C_l} = \sum_k p(C_l|\Psi_k)\Psi_k \tag{5.3}$$

In our implementation, we have chosen to use hard assignment of the SIPs to the clusters which renders the probability $p(C_l|\Psi_k)$ of assigning pattern Ψ_k to cluster C_l binary.

5.2.4 A Joint Model for Visual Categorization from Supervised, Semi-Supervised and Unsupervised Input

In this section, we present a model for visual category recognition that combines different levels of supervision to a joint model. The key ingredient is the scale-invariant pattern (SIP) representation from Section 5.2.1, which we use throughout.

Supervised Categorization To provide basic functionality for our system, we describe how supervised categorization is implemented. Similar to clustering in Section 5.2.3, we model each category A_i by a single representative $\bar{\Psi}^S_{A_i}$ (superscript S denotes the supervised model). This is done by summing over all training patterns available for that category

$$\bar{\Psi}^S_{A_i} = \sum_{j \in \mathbb{S}^{A_i}} \Psi_j, \tag{5.4}$$

where $\mathbb{S}^{A_i}$ denotes the indices of the SIPs that are labeled with category A_i in a supervised manner (e.g. “This is a bottle”).

Incorporating Semi-Supervision and Unsupervised Information in one consistent Framework We formulate the fusion of information obtained from supervised to unsupervised sources as an extension of the supervised case by assuming uncertainty about the correct labeling of the clusters C_l and their representatives $\bar{\Psi}_{C_l}$ from the unsupervised visual grouping step (Sec. 5.2.3):

$$\bar{\Psi}_{A_i} = \underbrace{\bar{\Psi}^S_{A_i}}_{\text{supervised}} + \underbrace{\sum_l p(A_i|C_l) \overbrace{\bar{\Psi}_{C_l}}^{\textit{unsupervised}}}_{\text{semi-supervised}} \tag{5.5}$$

$p(A_i|C_j)$ encodes the belief that cluster C_l contains instances of category A_i. How this probability is computed from a few interactions and updated by associating spatial expression with visual observations is described in Section 5.4.

To perform classification in the supervised and semi-supervised case, we evaluate the proposed model $\bar{\Psi}_{A_i}$ as well as $\bar{\Psi}^S_{A_i}$ for an observed pattern Ψ by using histogram

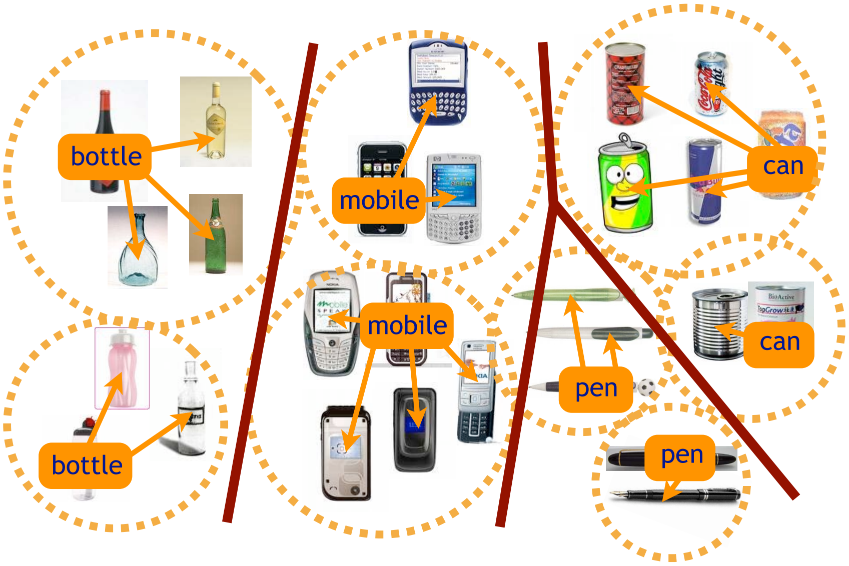


Figure 5.4: One basic approach to semi-supervised learning. Label information is propagated from labeled samples to unlabeled ones based on similarity of the observed samples. Note the refined decision boundary implied by this process.

intersection. Intuitively, the intersection measures to which percentage the model explains the observation, which we interpret as probability of belonging to the same class. In order to make models and observations comparable we normalize both to one. Bayes' rule is applied afterwards to obtain the model posterior:

$$p(\Psi|\bar{\Psi}_{A_i}) = \sum \min(\Psi, \bar{\Psi}_{A_i}) \tag{5.6}$$

$$p(\bar{\Psi}_{A_i}|\Psi) = \frac{p(\Psi|\bar{\Psi}_{A_i})p(\bar{\Psi}_{A_i})}{\sum_A p(\Psi|\bar{\Psi}_{A_i})p(\bar{\Psi}_{A_i})} \tag{5.7}$$

$p(\bar{\Psi}_{A_i})$ is the category prior, which we assume to be uniform. We decide for the category label with the highest posterior:

$$\hat{A}_i = \arg\max_{A_i} p(\bar{\Psi}_{A_i}|\Psi) \tag{5.8}$$

5.3 Dialogue Sub-System

Human-assisted visual learning is a form of *socially guided machine learning* (Thomaz, 2006). A human tutor interacts with the system, describing aspects of the environment the system is to learn. In our case, the tutor provides descriptions of the current visual scene. Typically, the tutor interacts with the system using spoken dialogue. For larger evaluations, we have mostly used typed- or scripted input.

The system uses a dialogue sub-system to try and comprehend what the tutor just said. This dialogue sub-system constructs a representation of the possible meaning(s) of an utterance, and then connects this representation to the larger dialogue context in which the utterance occurs. This makes it possible for the tutor to gradually provide information to the system, rather than all at once. The system keeps track of what objects have been introduced over the course of a dialogue, so the tutor can easily refer back to things already talked about: “This is a bottle. And there is a mobile. It is to the left of the apple. The bottle is to the right of the apple.” The system resolves pronouns (“it”) and anaphoric expressions (“the bottle”), identifying to which objects the provided information should be applied. (This sets our system somewhat apart from other approaches to relating language to the world, e.g. Roy (2002), Thomaz (2006), or Steels (2008), which all operate with individual utterances.)

The dialogue system constructs a representation of the possible meaning(s) for each utterance the tutor provides. The dialogue system uses a state-of-the-art approach to recognize spoken dialogue, using information about the current visual scene to prime speech recognition (Lison and Kruijff, 2008). Speech recognition yields a word lattice, representing probabilistically ranked possible sequences of words. The system subsequently parses the entire word lattice, using a Combinatory Categorial Grammar (Baldridge and Kruijff, 2003) parser[1]. The parser uses a CCG grammar to relate a possible syntactic structure for utterance, to the propositional meaning this structure expresses. As the parser processes a word lattice, representing different possibilities of what the system may have just heard, it needs to deal with the fact that there are potentially many analyses – and corresponding meanings. To this end, the parser uses discriminative statistical models to rank the analyses, picking out the most likely one in the given context (Lison and Kruijff, 2009). Altogether, this yields a robust way to overcome the typical problems of spoken dialogue processing, e.g. incomplete or incorrect input.

The parser provides a representation of utterance meaning as an ontologically richly sorted, relational structure similar to a description logic formula (Baldridge and Kruijff, 2001). After the dialogue system has interpreted the utterance meaning against the dialogue context model, it connects the resulting “contextual” meaning to information about the visual scene. This connection uses ontologies to mediate between linguistically expressed meaning, and the categories formed in the visual system (Jacobsson *et al.*, 2008). Using the hierarchical structure of ontologies, and the possibility to perform ontological inference over instances on these ontologies, provides a more general and better scalable approach to ”visual grounding” of language than provided by the string-based approach proposed in e.g. Roy (2002), or previous ontology-based approaches such as Kruijff *et al.* (2006a,b).

In our scenario, utterances are typically predicative copulative sentences in indicative mood (i.e. ”X is Y”), which assert that a given predication (”Y”) holds for the subject of the sentence (”X”). In our examples, the predication consists of a phrase that encodes a spatial relation (e.g. ”left of the bottle” or ”below the apple”). In the logical form, the subject is represented as the <Restr> of the state

---

[1] `http://openccg.sourceforge.net`

description that is denoted by the utterance, whereas the predication is represented as <Scope>.

We can thus easily derive the spatial configuration asserted in an utterance from its logical form representation (cf. also Kelleher *et al.* (2006a)). The following example shows such a logical form that is the result of the parsing process of the utterance ”the mobile is left of the bottle”:

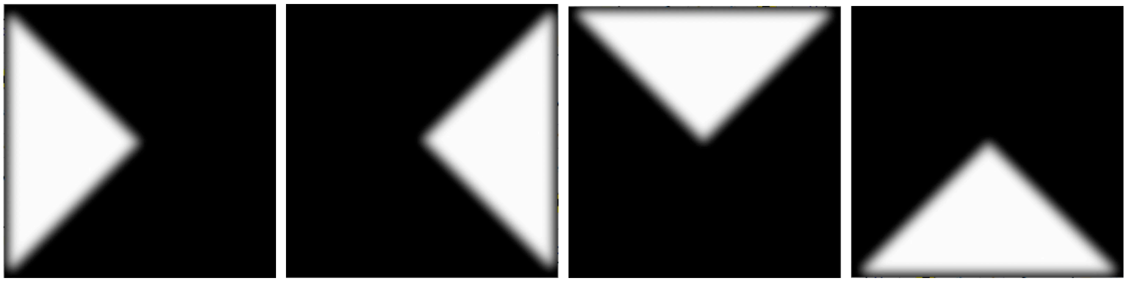
```
@b1:state(be ^
          <Mood>ind ^
          <Restr>(m1:thing ^ mobile ^
                  <Delimitation>unique ^
                  <Number>sg ^
                  <Quantification>specific_singular) ^
          <Scope>(l1:region ^ left ^
                  <Plane>horizontal ^
                  <Positioning>static ^
                  <Dir:Anchor>(b2:thing ^ bottle ^
                               <Delimitation>unique ^
                               <Number>sg ^
                               <Quantification>specific_singular)))
```

The logical form provides detailed information (at a linguistic level) about what the tutor just said about objects, events, and the relations between them. For an object, it represents an identifier and its ontological type ($m1 : thing$) and a proposition ($mobile$). The identifier is unique throughout the dialogue, and through reference resolution can help to relate mentions of an object. In addition, the logical form specifies information such as delimitation and quantification: how many objects we are talking about ($Quantification$), and to what extent they are easily identifiable in the visual scene. This information aids first of all in reference resolution, and later on in resolving linguistic references to visual objects. The distinction between “restrictor” and “scope” identifies the predication relation: The tutor presupposes that the system can identify the mobile (old information; restrictor), and then asserts the *new* information that it is to the left of the bottle (scope). Making this fine-grained differentiations in what status a piece of information actually has (old? new?) makes it possible for the system to decide how best to use the information in the incremental learning process (cf. also Hawes *et al.* (2009)).

For a more detailed discussion of the dialogue system, we refer the interested reader to Kruijff *et al.* (2009).

5.4 Spatial Reasoning and Cross-Modal Association

Modeling spatial relations as perceived by the human is a challenge in itself, as issues like reference frame and context have to be handled appropriately in situated



(a) (b) (c) (d)

Figure 5.5: 4 non-parametric probability density functions for modeling the spatial relations $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$ between feature patterns in 2d image coordinates, where R corresponds to one of the spatial relations: (a) left of (b) right of (c) above (d) below

dialogue systems (Kelleher *et al.*, 2006b). Considering the scenarios and main focus of this chapter, we restricted ourselves to modeling four basic spatial relations $R \in \{\text{"leftof"}, \text{"rightof"}, \text{"above"}, \text{"below"}\}$. We employ triangular shaped distributions $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$ defined in 2d image coordinates, where objects are referenced by their patterns Ψ_i and $\text{pos}(\Psi_i)$ denotes their position in image coordinates. Although these distributions are represented as non-parametric kernel densities which lend themselves to online updating, we don't explore this option here and keep them fixed in the experiments. Figure 5.5 visualizes the 4 distributions we are using.

Spatial Reasoning. We formulate the association of a spatial expression E extracted from an utterance (see Sec. 5.3) with two patterns Ψ_i and Ψ_j with positions $\text{pos}(\Psi_i)$ and $\text{pos}(\Psi_j)$ observed in scene S_k, as finding the most likely pair $\hat{P}_{i,j}$ of patterns: $\hat{P}_{i,j}^{(k)} = \arg\max_{P_{i,j}} p(P_{i,j}|E, S_k)$,where

$$\begin{aligned} p(P_{i,j}|E, S_k) &= p(\Psi_i, \Psi_j, \text{pos}(\Psi_i), \text{pos}(\Psi_j)|E, S_k) \\ &= p(\Psi_i|E, S_k)\; p(\Psi_j|E, S_k)\; p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|E, S_k), \end{aligned} \tag{5.9}$$

with

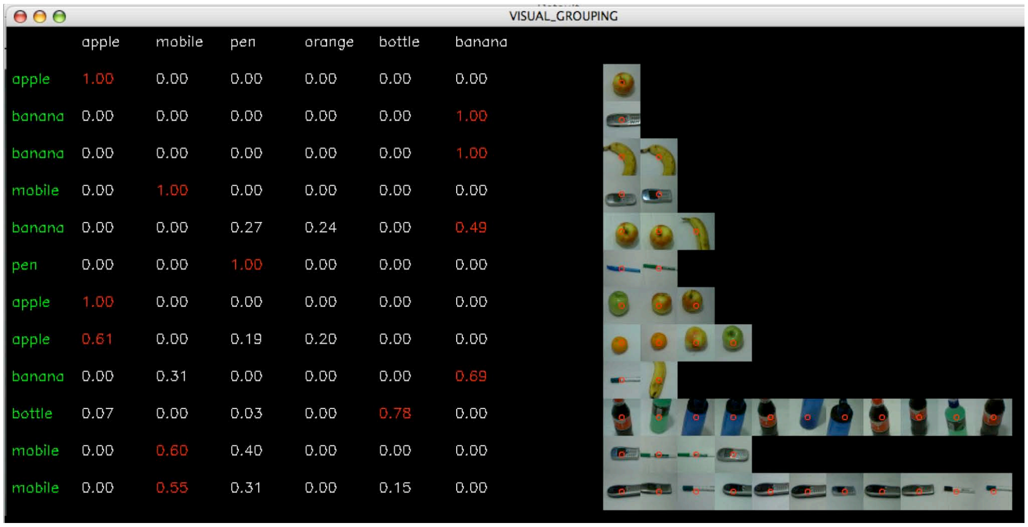
$$p(\Psi|E, S_k) = \sum_h p(\Psi|A_h) p(A_h|E, S_k). \tag{5.10}$$

We don't model a complete category system here, leave out contextual effects and assume certainty about the expression E referring to the categories A_{e_1} and A_{e_2} and the relation R. Consequently, the equation simplifies to

$$p(P_{i,j}|E, S_k) = p(\Psi_i|A_{e_1}) p(\Psi_j|A_{e_2}) p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R) \tag{5.11}$$

Finally, we insert the visual model from Eq. 5.5 to obtain a computational model:

$$p(P_{i,j}|E, S_k) = p(\Psi_i|\bar{\Psi}_{e_1}) p(\Psi_j|\bar{\Psi}_{e_2}) p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R) \tag{5.12}$$



	apple	mobile	pen	orange	bottle	banana
apple	1.00	0.00	0.00	0.00	0.00	0.00
banana	0.00	0.00	0.00	0.00	0.00	1.00
banana	0.00	0.00	0.00	0.00	0.00	1.00
mobile	0.00	1.00	0.00	0.00	0.00	0.00
banana	0.00	0.00	0.27	0.24	0.00	0.49
pen	0.00	0.00	1.00	0.00	0.00	0.00
apple	1.00	0.00	0.00	0.00	0.00	0.00
apple	0.61	0.00	0.19	0.20	0.00	0.00
banana	0.00	0.31	0.00	0.00	0.00	0.69
bottle	0.07	0.00	0.03	0.00	0.78	0.00
mobile	0.00	0.60	0.40	0.00	0.00	0.00
mobile	0.00	0.55	0.31	0.00	0.15	0.00

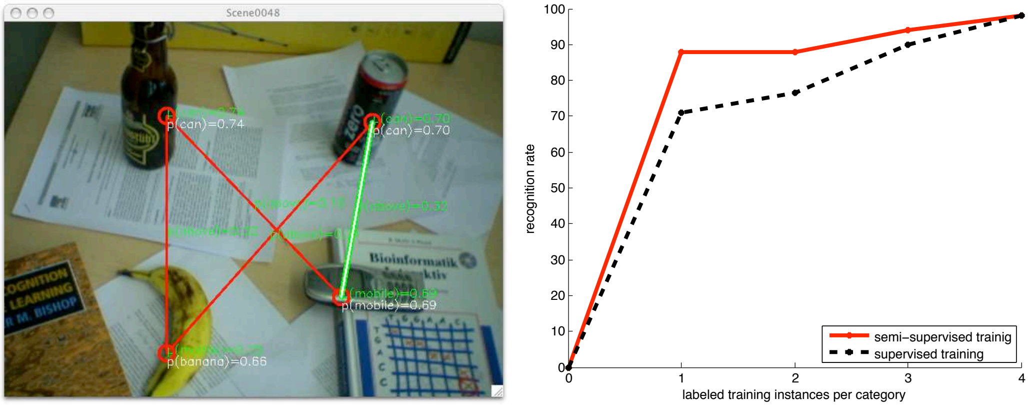
Figure 5.6: Semi-supervised learning is implemented in our system by assigning labels to the clusters only with a certain belief that represents the uncertainty resulting from the dialogue. Hereby, the system can recover from erroneous beliefs. The columns in the presented probability table represented the categories communicated via the language system and the rows correspond to the clusters obtained by the visual grouping step. The cluster members are visualized on the right.

This formulation facilitates incorporating information and belief from previous interactions as well as learning from scratch. If no information about the visual categories is available $p(\Psi_i|\bar{\Psi}_{e_1})$ and $p(\Psi_j|\bar{\Psi}_{e_2})$ become uninformative and the system relies only on its notion of spatial relations $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$. This can lead to wrong associations. In Section 5.5 we present an example and show that the system can successfully deal with this issue.

Cluster Labeling. We want to make use of the belief about associations between spatial expressions (Eq. 5.9) and objects in the scene to improve the label assignment $p(A_i|C_l)$ of the object clusters in Eq. 5.5. Therefore we accumulate the evidence for cluster C_l being labeled as category A_i by a simple count statistic $p(C_l|A_i)$ based on the maximum likelihood estimates of Equation 5.9. The probability for assigning label A_i to cluster C_l is obtained by applying Bayes' rule

$$p(A_i|C_l) = \frac{p(C_l|A_i)p(A_i)}{\sum_i p(C_l|A_i)p(A_i)}, \tag{5.13}$$

where we assume a uniform category prior $p(A_i)$. This closes the loop in our system as outlined in Fig. 5.1.



(a)

(b)

Figure 5.7: (a) shows an example scenario for propagation of labels from known categories to unknown ones. (b) shows the improvement of the semi-supervised approach over the purely supervised approach by exploiting information from unlabeled data.

5.5 Experiments

In the first part of our experiments, we describe two scenarios, that show the capabilities of our system to propagate information, resolve ambiguities and recover from errors. In the second part we perform a quantitative analysis to show that the unsupervised visual grouping step improves learning speed and accuracy with respect to the amount of provided supervision. Finally, we provide computation times for the individual modules and discuss the real-time capabilities of the system.

5.5.1 Label Propagation and Conflict Resolution

Scenario 1 - Label forward propagation. In the first scenario, one annotated example each for banana and mobile is presented to the system. Then the system observes the scene as shown in the screenshot in Figure 5.7(a) and the utterance "the can is above the mobile" is parsed. The red lines visualize the observed relations between objects in the scene. Unlikely ones have already been pruned away by the system. By generalizing across category instances, the system identifies "mobile" and "banana" correctly (with probabilities 0.69 and 0.66 respectively) while evaluating the "mobile" model for the banana results in a low probability of 0.15. Consequently the most likely relation is inferred correctly and displayed in light green. A model for the category "can" is created and the observed mobile is added to the existing model for "mobile". In fact, the figure shows the state in which the acquired "can" model is already used for detection. The can is detected correctly, but also the bottle gets a high score for the "can" model, as it's the best explanation given the learned categories (banana, mobile, can).

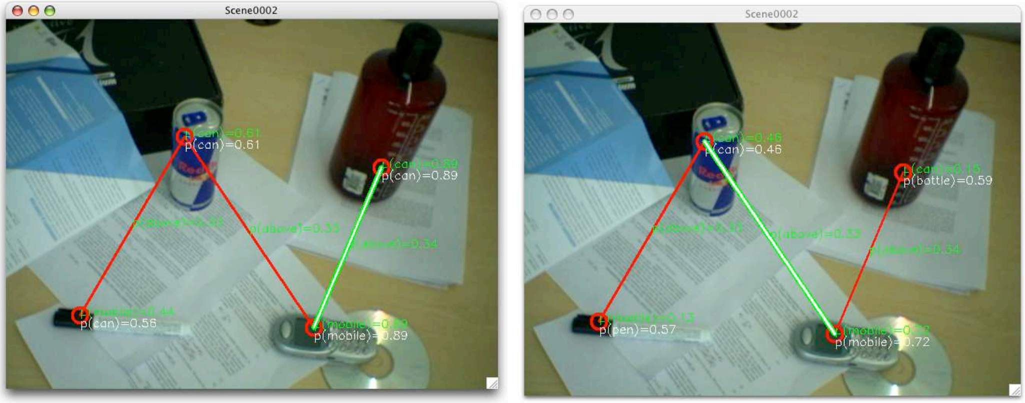


Figure 5.8: Scenario which shows how the system updates associations (light green) to recover form an incorrect belief.

Scenario 2 - Label backward propagation. In the second scenario, we show how the system can recover from erroneous beliefs and update its models accordingly. The system starts without any knowledge about visual categories. Figure 5.8 shows a screenshot displaying the scene as observed by the system, which is accompanied by the utterance "the can is above the mobile". Using the same visualization as in the previous scenario, it can be seen in the left image that the most likely relation inferred by the system is wrong. Now we provide the system with supervised knowledge of the visual categories bottle and pen. Revisiting the scene in memory, the object probabilities get updated and the belief about the associated relation gets changed to the correct one as shown in the right image.

5.5.2 Quantitative Evaluation

Semi-Supervised Learning. We perform a quantitative analysis by taking 2 images of 5 instances for each of the categories: mobile phone, pen, bottle, can and apple. We gradually increase the training set from one instance for each category to four instances. Figure 5.7(b) shows that the semi-supervised learning (Sec. 5.2.4) outperforms the purely supervised learning, as the few available labels get propagated to the unlabeled data (Eq. 5.13) which was clustered by the visual grouping step (Sec. 5.2.3). The experiments were performed using 5 fold cross-validation.

Speed. The system as described in this chapter runs at about 1Hz on a CoreDuo 2GHz laptop when detection and categorization are performed. An update of the clustering and spatial reasoning takes about 2 seconds total. Therefore the system is fast enough to operate interactively with a human tutor.

5.6 Conclusions

This chapter presents a system for cross-modal learning that combines unsupervised and supervised information in a unified framework. The mechanism that associates expressions from language with the visual input can resolve ambiguous input and recover from erroneous beliefs. The experimental section provides qualitative as well as quantitative results that show these capabilities of the system. Finally, we were able to cut down the computing time to a level at which a human can interact with the system as a tutor.

6

Decomposition of Visual Categories Using Topic Models

Object representations for categorization tasks should be applicable for a wide range of objects, scaleable to handle large numbers of object classes, and at the same time learnable from a few training samples. While such a scalable representation is still illusive today, it has been argued that such a representation should have at least the following properties: it should enable sharing of features (Torralba *et al.*, 2007), it should combine generative models with discriminative models (Jaakkola and Haussler, 1999; Fritz *et al.*, 2005) and it should combine both local and global as well as appearance- and shape-based features (Leibe *et al.*, 2005). Additionally, we argue that such object representations should be applicable both for unsupervised learning (e.g. visual object discovery) as well as supervised training (e.g. object detection). Therefore, we aim to combine in this chapter our previous efforts of hybrid modeling from Chapter 3 with ideas of combining different levels of supervision from Chapters 4 and 5 to obtain an approach that shows flexibility and adaptivity along all 3 axes: Modeling, Representing and Learning.

We present a novel method for the discovery and detection of visual object categories based on decompositions using topic models. The approach is capable of learning a compact and low dimensional representation for multiple visual categories from multiple view points without labeling of the training instances. The learnt object components range from local structures over line segments to global silhouette-like descriptions. This representation can be used to discover object categories in a totally unsupervised fashion. Furthermore we employ the representation as the basis for building a supervised multi-category detection system making efficient use of training examples and outperforming pure features-based representations. The proposed speed-ups make the system scale to large databases. Experiments on three databases show that the approach improves the state-of-the-art in unsupervised learning as well as supervised detection. In particular we improve the state-of-the-art on the challenging PASCAL'06 multi-class detection tasks for several categories.

The chapter is structured as follows. In Section 6.1 we describe how the generative decomposition is learned from data. The obtained representation is used in Section 6.2 for unsupervised learning problems whereas Section 6.3 builds a full object class detection system on top of it. Finally Section 6.4 provides further quantitative evaluations of the model and a comparison to the state-of-the-art on the challenging PASCAL'06 database as well as on the ETH shape database.

6.1 Decomposition of Visual Categories

In this section we describe our approach to the decomposition of multiple visual categories by combining dense gradient representations and topic models. Starting from the image, we first present our data representation. Then we describe how we apply the topic model to this representation and provide visualizations and insights for the obtained model as well as a quantitative evaluation on an unsupervised learning task.

6.1.1 Data Representation

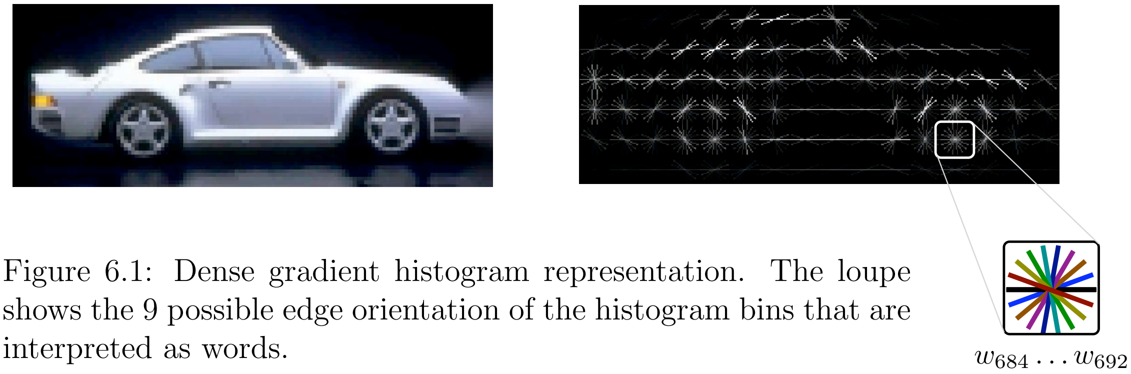


Figure 6.1: Dense gradient histogram representation. The loupe shows the 9 possible edge orientation of the histogram bins that are interpreted as words.

Inspired by Dalal and Triggs (2005), we compute gradients on each color channel of the images and use the maximum response (at each pixel) to obtain a grid of histograms that overlays the image. Each histogram in the grid has 9 orientation bins equally spaced from 0° to 180° to represent the unsigned gradient orientation. An example of such an encoding is visualized in Figure 6.1. In each cell, the 9 possible edge orientations associated with the orientation bins are displayed by short lines. The grayscale value encodes the accumulated gradient magnitude in each bin. The size of the cells in the grid is 8×8 pixels.

As the following topic models operate on discrete word counts, we normalize the histograms to have a constant sum of discrete entries. We decided not to compute a redundant coding like the blocks in the HOG descriptor of Dalal and Triggs (2005) as we believe that the introduced non-linearities by local normalization would hinder the fitting of the probabilistic model.

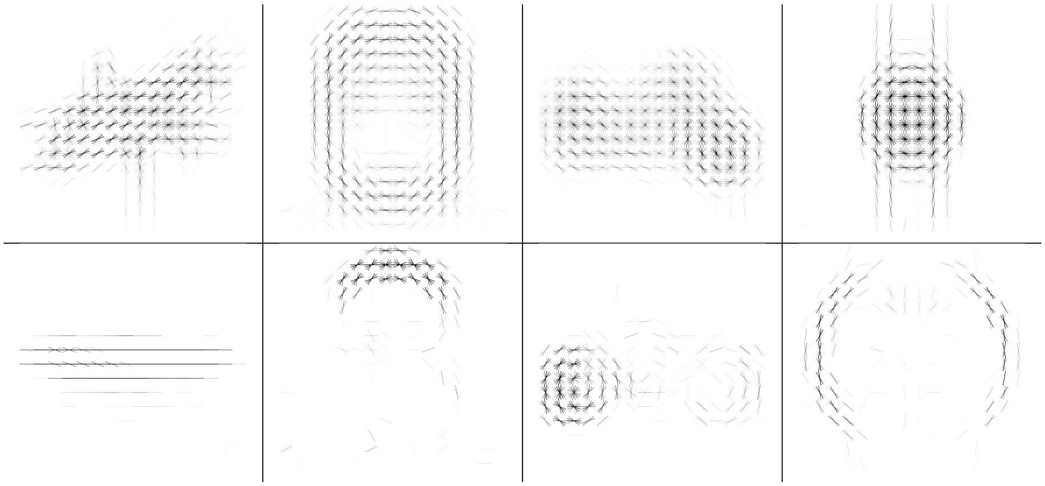


Figure 6.2: First row: example topics of 8 topic model for classes airplane, face, motorbike, watch. Second row: example topics of 50 topic model for the same classes.

6.2 Discovery of Visual Categories

In this section we describe how the representation from Section 6.1.1 is linked to the probabilistic topic model (see Section 2.2.3) and perform a quantitative evaluation on an unsupervised learning task.

We use the orientation bins of the histograms described in Section 6.1.1 as word vocabulary in Section 2.2.3. For histograms computed on a m by n grid with b bins for each orientation histogram, our vocabulary is of size $|V| = m \cdot n \cdot b$. As each word is associated with a gradient orientation at a grid location, this representation preserves quantized spatial information of the original gradients. The topic model is trained on the documents given the encoded training examples, as outlined in Section 2.2.3. The representations that we promote are given by the topic distribution $\theta^{(d)}$ of the document in the latent space. We will refer to them also as the topic activations.

To prove the effectiveness of our representations and to compare our work with previous approaches we first present quantitative results on the unsupervised ranking task of Fergus *et al.* (2005a) and then provide further insights connected to the multi-class data we use in Section 6.4.3.

6.2.1 Unsupervised Google Re-Ranking Task

Previously, Sivic *et al.* (2005) used topic models on local feature representations for unsupervised learning. Fergus *et al.* (2005a) extended their approach to encode spatial information. As the latter can be seen as the sparse counterpart to our dense representation, we compare on the unsupervised image re-ranking task specified in Fergus *et al.* (2005a). The provided data sets are results of image google queries. The task is to re-rank the images so that relevant ones appear first. The main

	airplane	cars rear	face	guitar	leopard	motorbike	wrist watch	average
out method	100%	83%	100 %	91%	65%	97%	100%	91%
Fergus Fergus *et al.* (2005a)	57 %	77%	82%	50%	59%	72%	88%	69%
Schroff Schroff *et al.* (2007)	35%	–	–	29%	50%	63%	93%	54%

Table 6.1: Comparison to other approaches on re-ranking task of google images. Performance is measured in precision at 15% recall. In contrast to the other methods our approach does not use any validation set.

challenge is to extract the relevant information which is hidden in an image set containing up to 70% junk images in an unsupervised fashion. Given that our representation effectively encodes the object structures, we expect our data to live in compact subspaces of the latent space. Therefore, we perform k-means clustering on the activations and consecutively accept the clusters with the most samples. The precision we obtain in this manner at 15% recall is shown in Table 6.1 and compared to our competitors. The average precision of 69% obtained by Fergus *et al.* (2005a) and 54% obtained by Schroff *et al.* (2007) is surpassed by our approach which obtains an average precision of 91%. This performance is obtained without using the provided validation set which the other two approaches use. Although our method performs worst on the leopard data, we still improve over Fergus *et al.* (2005a). This is surprising as one would have suspected, that the local feature-based approach is more suited to encode the spotted texture of these animals. We account the success of our method to the added expressiveness by enabling the discovery of reoccurring contour fragments and edge segment like structures. Due to the dense and localized nature of our input features, we are more flexible to adapt to the object outline and to neglect background information. Figure 6.2 shows some topics from the presented experiment that expose these characteristics. Furthermore, in contrast to local feature-based methods our representation can easily be visualized (see Figure 6.2), which lend itself also to interaction and inspection by a user.

6.2.2 Unsupervised Object Class Discovery

To extend our findings to the detection task that we are aiming for in Section 6.4.3, we extract our representation on the multi-category, multi-view PASCAL'06 dataset Everingham *et al.* (2006), in order to obtain a decomposition that is shared across categories.

In the first row of Figure 6.3 13 of 100 topics are visualized that were trained on the bounding box annotations of the training and validation data of the PASCAL'06 challenge. The rows below display the examples that activated this particular topic most. By activations, we refer again to the probability of the topics z given a specific document d, which we infer in the topic model. We observe that the topics capture different levels of object structure, ranging from global silhouettes (car rear in column 10 and side view in column 13) over localized parts (legs in column 3, bicycle frame in column 8 and bicycle wheels in column 12) to line segments and corners (corner in column 1 and line segments in column 2 and 4) . The model discovers distinctive parts that even separate several examples of different categories and their viewpoints although no such information was available to the system during training. Importantly, we can see that other topics like those that got activated on legs are shared across several categories, which is a desirable property of a compact decomposition in order to be scalable (Torralba *et al.*, 2007).

To illustrate that this is indeed an appropriate and effective approach to capture the variety of the data and to stress the power of modeling combinations of these discovered topics, we cluster the topic distributions as proposed in the last paragraph. Figure 6.4 shows in each row the 10 cluster members that are closest to the cluster center all of the 50 cluster centers. Keeping in mind that they are obtained in an entirely unsupervised fashion, the clusters turn out to be surprisingly clean.

We interpret these findings as strong evidence, that our model indeed captures an effective and low-dimensional representation for this difficult multi-category detection task.

6.3 Detection of Visual Categories

Based on the promising results on unsupervised learning in the last section, this section describes a complete system for supervised multi-category detection that leverages the learned representation.

6.3.1 Generative/Discriminative Training

As already argued in Chapter 3, the combinations of generative approaches with discriminative ones has shown to be very effective. The success of these combinations is based on their complementary strengths, that we have summarized in Section 2.1.1. We also exploit this idea in this chapter and complement the generative model described in Section 6.2 by a discriminative SVM classifier with an RBF kernel (Chang and Lin (2001)). In particular we train an SVM to discriminate between the topic distributions $\theta^{(d)}$ which are inferred for images containing the category of interest and others that do not contain these. By doing so, we seek to profit from the above mentioned benefits of the generative model combined with the discriminative classifier.

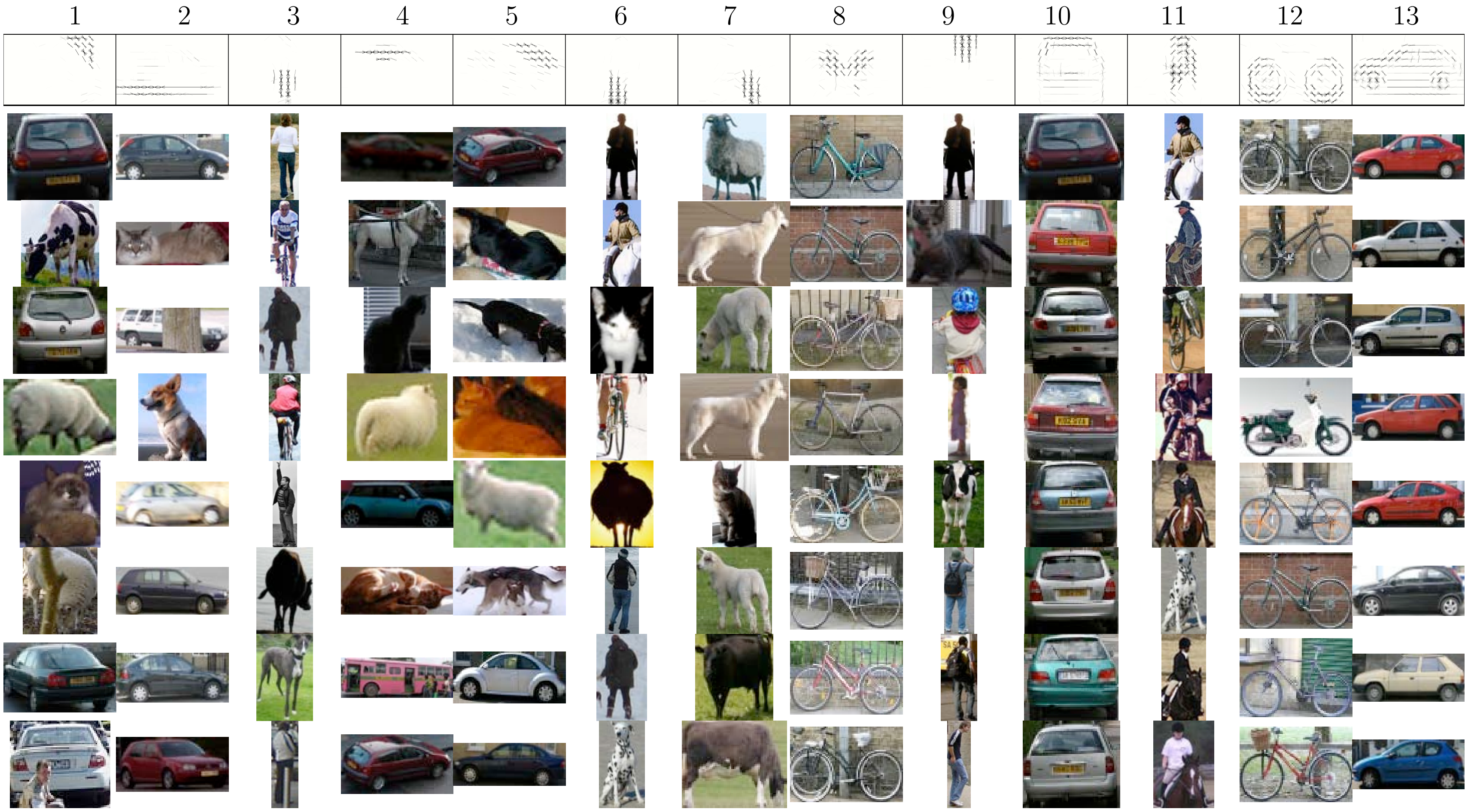


Figure 6.3: First row: example topics that were learned by the proposed approach across categories and viewpoints for the 10 classes of the PASCAL'06 data. Below first row: training images that activated the topic above most. The topics model local structures, line segments as well as silhouette-like structures. The topics are distinctive enough to separate several category members and even view-points. On the other hand they are general enough to be shared across categories and viewpoints.

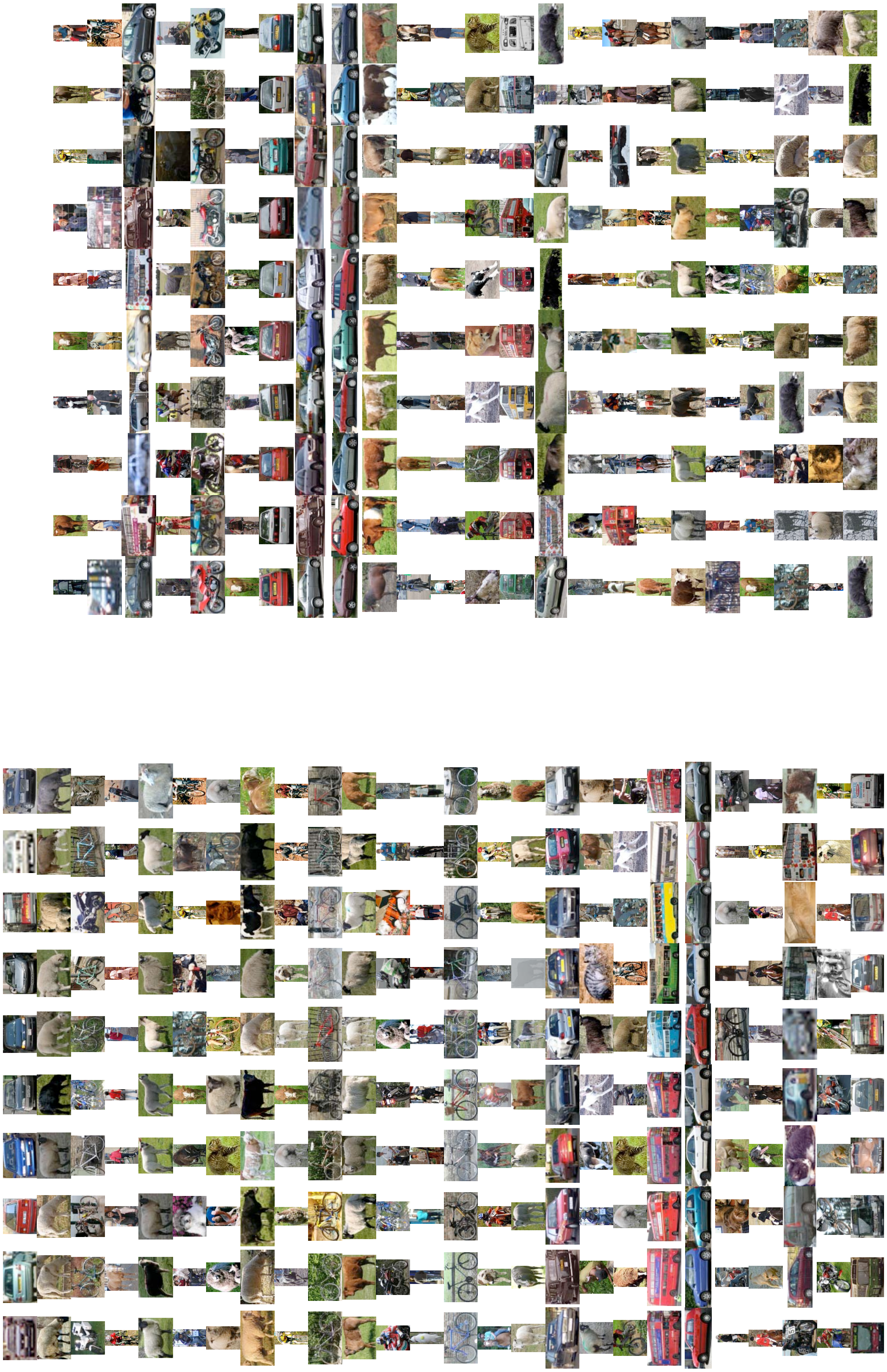


Figure 6.4: Unsupervised discovery of categories and viewpoints in PASCAL'06 data. The rows show for all 50 clusters those 10 examples that are closest to the cluster center. The left block visualizes the clusters 1 to 25 and the right block visualizes the clusters 26-50.

6.3.2 Sliding Window Approach to Detection

As proposed in Dalal and Triggs (2005) a sliding window approach can be done efficiently in this setting if the sliding window is always shifted by exactly one cell in x or y direction. In this case, the gradient histograms of the cell grid are computed once and for each sliding window the relevant sub grid is used.

Typically, sliding window techniques not only assign a high score for the correct location and scale in an image, but also for test windows that have a small offset in space and scale. We use a simple greedy scheme to cope with this issue: While there are unprocessed windows in an image, we accept the one with the highest score and reject all other windows that fulfill the symmetric overlap criterion

$$\max\left(\frac{A_i \cap A_j}{A_i}, \frac{A_i \cap A_j}{A_j}\right) > 0.3 \tag{6.1}$$

where A_i and A_j are the areas covered by the two windows. As the bounding box scores from our approach turn out to be surprisingly consistent over different scales, this basic scheme has proven to work well in our setting.

Of course multi-scale detection task ranging over multiple octaves requires the investigation of large number of test windows – typically more than 10000 per image. While feature extraction and SVM classification are fast, our approach requires inference in the topic model for each test window rendering the method computationally infeasible for applications of interest. Therefore, we dedicate the following section to describe speed-ups that make our approach applicable to large databases.

6.3.3 Speed-ups: Linear Topic Response and Early Rejection

While we use the Gibbs sampling method (Griffiths and Steyvers, 2004) to estimate the model, we use the variational inference method described in Blei *et al.* (2003b) for test as it turns out to be computational more efficient in our setting. For more substantial improvements, we propose to compute a linear topic response to get an initial estimate on the topic activations. The aim is to avoid the more expensive inference scheme by performing an early rejection of the test windows. Different to linear methods like PCA, where there is linear dependency between the feature space and the coefficient space, the mixture coefficients of the topic distribution have to be fitted to the observation. This means that each word/feature can be associated to different topics depending on its context (presence of other features) and therefore also lead to strengthening or inhibition of other topic activations. This requires an iterative technique to find the best reconstruction. Therefore we ask the question of how important this iterative fitting is and how much performance we loose by reverting to the following simple, linear approximation of the dependency between observed feature histogram x and topic activations $\theta^{(d)}$:

$$\tilde{\theta}^{(d)} = \left(\phi^{(1)} \ldots \phi^{(T)}\right)^t x, \tag{6.2}$$

In fact, our results on the UIUC single scale database show that there is a significant loss of about 8% in equal error rate performance (see Section 6.4.2), but a more detailed analysis on the UIUC multi-scale database reveals interesting results. Although, the linear approximation might be quite coarse, it can still be used for early rejection of test windows. It turns out, that full recall is achieved for the 2500 highest scored windows of a total of 2,826,783. As a consequence, more than 99.9% of the test windows can be handled by the linear computation that we measured to be 166 times faster than the proper inference. Taking all optimizations together we can cut down the computation time by a factor of 180 which corresponds to an reduction from one hour to around 20 seconds per image (AMD Opteron 270 (Dual-Core), 2.0 GHz).

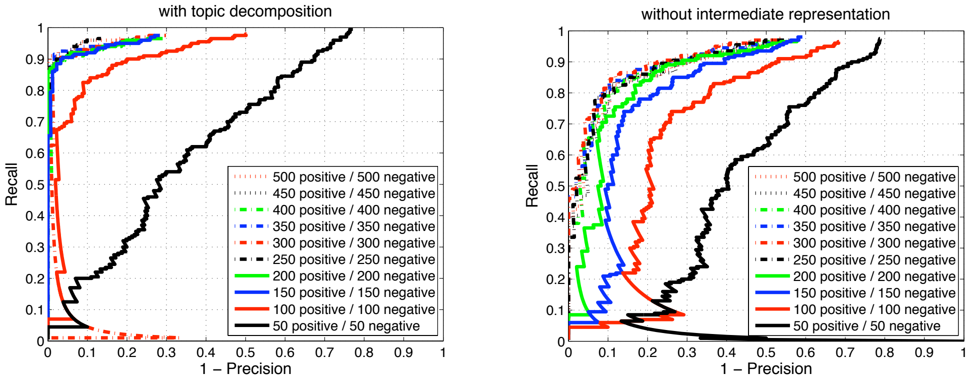
6.4 Experiments

This section is divided into 4 parts. First, we show that our approach makes efficient use of the provided training examples by comparing to a baseline experiment on the UIUC single scale car database. Second, we evaluate different methods for estimation of the topic model on the UIUC multi-scale database and compare the obtained performance to previous work. Third, we present results on the PASCAL challenge 2006 data, that outperform the state-of-the-art on three of the ten categories. Fourth, we compare to a shape based approach on the ETH shape database to underline the versatility and adaptivity of our approach.

6.4.1 Efficient Use of Training Examples and Parameter Selection

To select parameters appropriate to our problem domain, we run detection experiments on the UIUC single scale car database which consists of a training set of 550 car and 500 background images of small size, while the test set has 170 images showing side views of cars in street scenes at a fixed scale. It turns out that the heuristic specified in Steyvers and Griffiths (2007) for selecting the hyperparameters α and β works very well for our setting. Therefore we use $\alpha = 50/\#topics$ and $\beta = 0.01$. We obtain best performance using 30 topics and a grid size of 16×6 for the gradient histograms.

To show that our approach makes efficient use of the provided training examples, we compare to a baseline experiment that does not use the proposed topic representation. Figures 6.5(a) and 6.5(b) show the precision-recall curves of our system, when trained on different numbers of positive and negative examples. We start with 50 car and 50 background images and increase by 50 until we use the full training dataset. The maximum performance is rapidly reached using only 150 positive and 150 negative examples. In contrast, the linear SVM trained on the same data representation but without our representation has a much slower learning curve. In fact the performance is 9.5% below the equal error rate of our new approach using



(a) (b)

Figure 6.5: (a) and (b): Comparison of learning curve for proposed intermediate representation versus SVM on pure features on UIUC single-scale database.

250 positive and 250 negative examples. We also tried RBF kernels, but obtained similar, inferior results.

We account this significant improvement to the generative properties of our model inferring a generative decomposition of the presented data. We conclude, that this low dimensional representation simplifies the learning problem for the discriminative SVM classifier, which leads to more efficient use of training examples.

6.4.2 Comparison of Methods for Estimation and Evaluation of Approximate Inference

In this section we test the model that we trained for the UIUC single scale database on the multi-scale version and compare different estimation schemes for the topic model during training (Blei *et al.*, 2003b; Griffiths and Steyvers, 2004). We also evaluate the linear topic activations for testing that we proposed in Section 6.3.3. The results are reported in Figure 6.6(a). The estimation method based on Gibbs sampling (Griffiths and Steyvers, 2004) leads to similar performance as the variational inference method (Blei *et al.*, 2003b), but shows better precision. We notice that the automatic selection of α that we use for the variational approach converged to a value of 0.373 which enforces less co-activation and therefore less sharing of topics. By visual inspection of the topic-distributions, we confirmed that the method of Blei *et al.* (2003b) learned more global topics, while the ones obtained by the Gibbs sampling method tends to be a little sparser. We believe that for detection tasks the second is to be preferred, as global representations can easier be mislead by effects like occlusion, as it is also supported by our results.

Replacing the proper inference by the linear approximation (Section 6.3.3) results in the third curve displayed in Figure 6.6(a). This confirms the importance and

bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
49.75%	25.83%	50.07%	9.09%	15.63%	4.55%	9.40%	27.43%	0.98%	17.22%

Table 6.2: Average precision achieved on the PASCAL'06 database.

superiority of the proper inference in comparison to linear topic activations. For this comparison we use non-maxima suppression in combination with the linear approximation scheme while it is switched off when used for early rejection to achieve maximum recall.

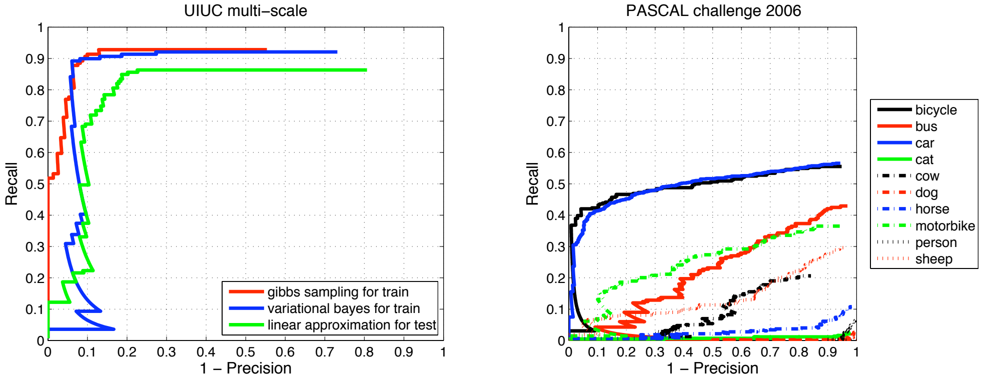
The best result obtained by the Gibbs sampling approach with an equal error performance of 90.6% outperforms the results we presented in Chapter 3 and are on par with the result of Mutch and Lowe (2006). The best performance on this dataset have been reported by Wu and Nevatia (2007) with 93.5% and Mikolajczyk *et al.* (2006) with 94.7%, where the later used a different training set.

6.4.3 Comparison to state-of-the-art on PASCAL'06 VOC detection challenge

We evaluate our approach on the competition 3 of the PASCAL challenge 2006 Everingham *et al.* (2006) that poses a difficult detection problem as 10 visual categories are to be detected from multiple viewpoints over a large scale range.

We leave the hyperparameters untouched, but increase the number of topics to 100 and adopt the aspect ratio of the grid to 16×10. To reduce confusion between categories and the number of false positives, we adapt a bootstrapping strategy. First we train an initial model for each category versus the other categories. This model is then used to generate false positives on the training set (see also Osuna *et al.* (1997); Fritz *et al.* (2005); Dalal and Triggs (2005)). Up to 500 of the strongest false detection are added for each detector to its training set and the model is retrained. The average precisions of the final detector of all 10 categories on the test set are shown in Table 6.2 and the corresponding precision-recall curves are plotted in Figure 6.6(b). Figure 6.7 shows some example detections of the system.

We outperform all other competitors in the 3 categories bicycle, bus and car by improving the state-of-the-art (Everingham *et al.* (2006)) on this dataset by 5.75%, 9.14% and 5.67% in average precision respectively. In particular we surpass the fully global approach Dalal and Triggs (2005) that our method was motived by. Compared to Chum and Zisserman (2007) we improve on bicycles and bus only by 0.65% and 0.93%, but again significantly on cars with 8.87%. However, in contrast to Chum and Zisserman (2007) we do not use the viewpoint annotations to train our approach. For the other categories, we perform about average, but also showed some inferior results on the highly articulated categories. We are currently investigating means to make the approach less rigid and carry over the good results from the first 3 categories to the other ones.



(a) (b)

Figure 6.6: (a) Performance on UIUC multi-scale dataset using topic model estimated via Gibbs sampling vs variational bayes approach compared to using pseudo topic activations. (b)Precision-Recall curves on the PASCAL VOC challenge 2006. Precision-Recall curves and example detections.

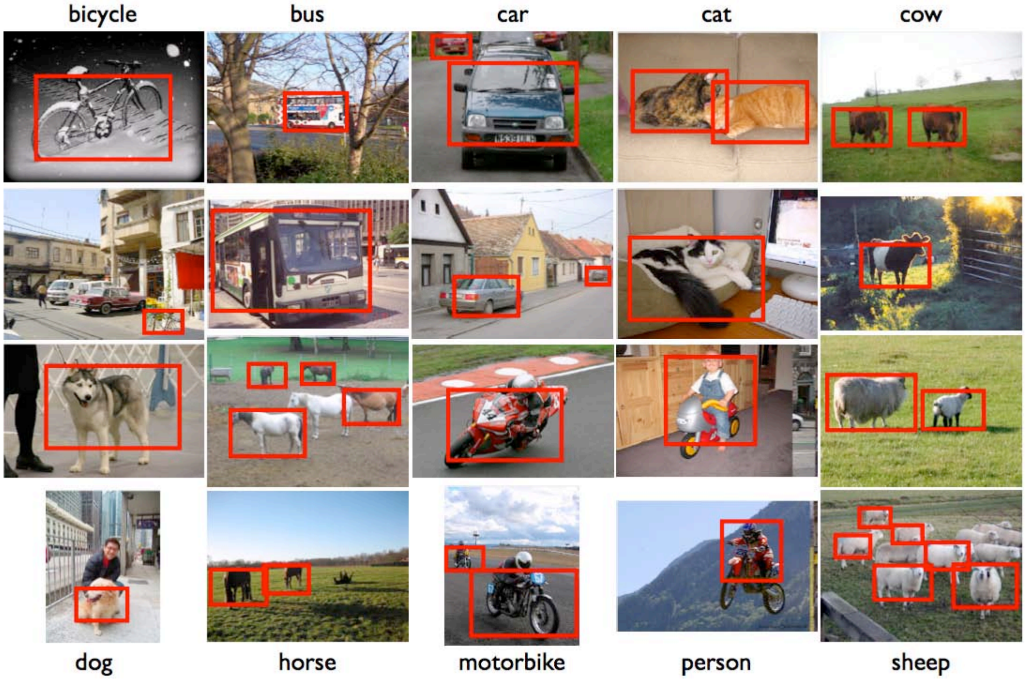


Figure 6.7: Example detections on the PASCAL VOC challenge 2006.

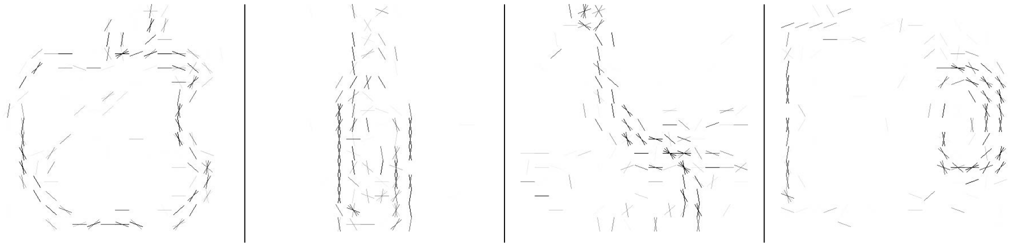


Figure 6.8: Example topics of 100 topic model jointly learned on apple-logos, bottles, giraffes, mugs and swans.

	Applelogos	Bottles	Giraffes	Mugs	Swans	average
our method	89.9%(4.5)	76.8%(6.1)	90.5 %(5.4)	82.7%(5.1)	84.0%(8.4)	84.8%
Ferrari *et al.* (2007)	83.2%(1.7)	83.2%(7.5)	58.6 %(14.6)	83.6 %(8.6)	75.4 %(13.4)	76.8%

Table 6.3: Comparison against shape-based approach of Ferrari *et al.* (2007) on ETH shape database. Average detection-rate at 0.4 false positives per image averaged over 5-folds. Standard deviation is specified in brackets.

6.4.4 Comparison to shape features on ETH shape database

As pointed out in the previous experiments, our representation learns features with different characteristics from local to global and is in particular also capable of modeling contours. Therefore, we ask the question how our representation compares to shape-based approaches. We compare to Ferrari *et al.* (2007) on the ETH shape database using the same detection system with the same settings as described in the last section. Example topics that were learnt across the 5 classes are depicted in Figure 6.8. Note how the more rigid shapes like apple logo and bottle are represented as a whole, while the topics for giraffe and mug focus on the more stable parts like the back of the giraffe and the combination of handle and sidewall of the mug. Using five fold cross-validation as proposed in Ferrari *et al.* (2007), we obtain the results presented in Table 6.3. Averaged over all classes we improve the performance of Ferrari *et al.* (2007) by 8.0% to 84.8%. On apple logos, giraffes and swans, we improve the performance by 6.7%, 31.9% and 8.6% respectively. On mugs our approach performs comparable and on bottles it looses 6.4%. We account the worse performance on the bottles to the shape which is less discriminant with respect to the background. Some example detections are shown in Figure 6.9. Note how the objects are well localized and the approach even detects a half visible mug that is not annotated (third image, top row). As the database was designed to test shape-based approaches, the improvements obtained by our approach underlines the versatility and adaptivity of the learnt representation.

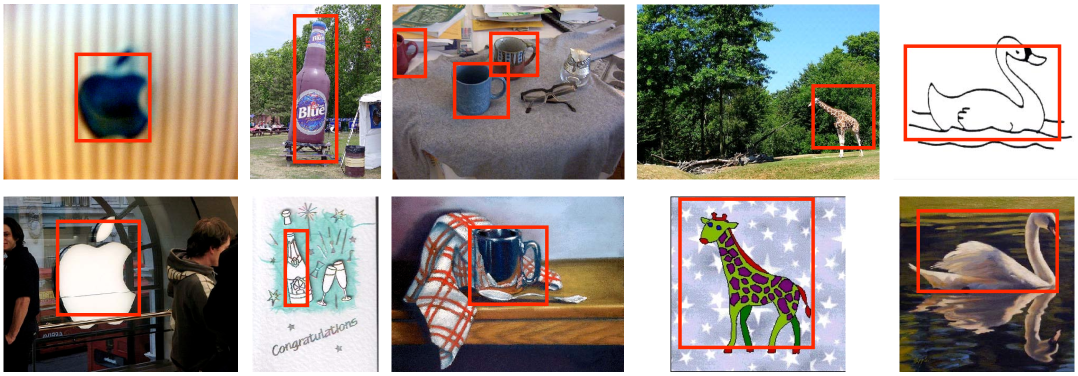


Figure 6.9: Example detections on the ETH shape database.

6.5 Conclusions

We present a novel method for representing multiple categories from multiple viewpoints and successfully employ it in various settings ranging from unsupervised learning to supervised detection tasks. In various experiments our approach shows superior performance with respect to purely local, shape-based or global approaches. Our representation has proven effective yet also efficient in showing an increased learning curve in the detection setting. Beyond the modeling aspects, we pay particular attention to computational feasibility that enables scalability to large databases. Lastly, we want to highlight the results on the challenging PASCAL'06 dataset where we improve the state-of-the-art on three categories to underline our contribution to category modeling in the context of a complete detection system.

7

Extensions Towards Explicit Multi-View Modeling

This chapter builds on the generative decomposition developed in Chapter 6 and develops multiple extensions to deal with more challenging training and test data as presented in the PASCAL VOC challenge 2007 (Everingham *et al.*, 2007). Particular attention is paid to the corrupted training data and multi-view detection.

Therefore, Section 7.1 proposes an effective way to sort training examples. We use the document likelihood given a trained generative decomposition based on and LDA model as described in Chapter 6 as a score. The ranking is then used to focus on the more prototypical examples that have less artifacts like heavy occlusion or poor illumination.

Although the approach presented in Chapter 6 can already deal with a decent amount of viewpoint variation, Section 7.2 sheds more light on how multi-viewpoints are represented and proposes extension to deal with these kind of variations in a more explicit way. After a short review of the key challenges encountered in multi-view models, we present experiments on unsupervised viewpoint recovery and aspect prediction for detection.

7.1 Data Cleaning by LDA Likelihood Ranking

Setting up large datasets is a non-trivial problem. The labeling requires tedious manual work and still inconsistencies and biases are unavoidable (Ponce *et al.*, 2006). One particularly lively debated issue is whether the set should provide rather clean and prototypical examples (e.g. Yao *et al.* (2007)) or if the samples should come with all the corruptions typical for real-world settings (e.g. Everingham *et al.* (2007)). Both approaches have their own appeal and inherent limitations and the arising conflict on how to setup a database properly is illustrated by the following examples.

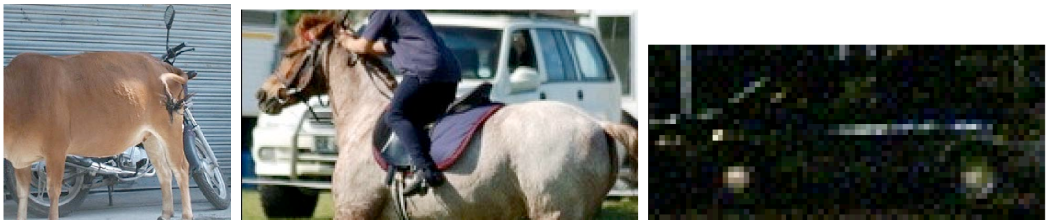
Machine learning methods have shown to be very effective when a sufficiently large sample of these corruptions is presented.For example, textures were successfully recognized despite scale and illumination changes (Hayman *et al.*, 2004) by providing examples of these artifacts. For other types of real-world artifacts, this approach is unlikely to scale. It is impossible to capture all kinds of possible occlusions in a database - even a representative sample seems illusive. Therefore, occlusion is one of the artifacts that is typically dealt with at test time. Still there might be some occluders that are so frequent that they provide contextual information and shouldn't be ignored.

PASCAL datasets contain lots of real-world artifacts and it is unclear which ones should be discarded or kept for training. Additionally, the associated object class detection challenge doesn't allow for hand-picking examples. However, the necessity of a preprocessing step becomes apparent when inspecting some of the training examples as depicted in Figure 7.1. The first two examples show object instances occluded by instances of other classes, which is likely to increase category confusion. The third example shows a very poorly illuminated object. It is at least questionable if such examples provide useful information for training.

This motivated us to develop a fully automatic data cleaning method that is *specific to our model*. The underlying assumption is that examples that contain little information or that are out of the focus of the method will hamper the training process and therefore should be excluded. For example, the data set doesn't provide a large statistic of rotated examples. Therefore, we want to realize this right at the beginning of the training process and improve the model fit on the rest of the data by sacrificing the rotated examples.

The solution we propose is to use the same generative decomposition we use for detection as in Chapter 6 and do an unsupervised re-ranking of the training images according to the likelihood of the employed LDA model. In order to obtain a stronger pruning effect we reduce the capacity of the model by lowering the number of topics to 5 topics and train for each category a separate model. All trained topics for all 20 classes of the PASCAL VOC challenges are depicted in Figure 7.2 and Figure 7.3. Some topics reveal unexpected regularities in the data. For examples the 4th topics of the horse category (Figure 7.3) shows a barrier as many horses are imaged during jumps during show jump events.

An example for the obtained ranking on the car category is shown in Figure 7.4. The top 20 examples on the top are rather clean but still capture a reasonable amount of variance. Also the fact that these best ranked examples come from very different scales (as can be seen by the resolution) is worth mentioning. On the contrary, the 20 lowest ranked examples on the bottom are mostly outliers in the sense that they are weakly represented in the database in terms of viewpoint or rotation, expose bad imaging conditions or heavy occlusion. For the challenge entry we decided for a rather clean set by talking the top 50% of the re-ranked training images.



(a) motorbike example corrupted by occlusion

(b) car example corrupted by occlusion

(c) car example badly light

Figure 7.1: PASCAL'07 training set contains examples that expose heavy occlusion and lighting artefaces.

7.2 Generative Object Decompositions for Multi-Viewpoint Modeling

Inferring the position and the orientation of a potentially occluded object in the presence of cluttered background is a challenging.

First, in Section 7.2.1, we briefly review different approaches to multi-viewpoint object class recognition. Then, we investigate the capabilities of modeling and discovering of the approach from Chapter 6 in more detail. Our efforts towards unsupervised viewpoint discovery are summarized in Section 7.2.2. This topic representation is then applied to a state-of-the-art benchmark dataset, the PASCAL VOC 2007 challenge. Our results in this competition confirms the applicability of our new model to the challenging problem of multi-view modeling. Section 7.3 summarizes the results.

7.2.1 Towards the Representation of Multi-View Object Categories

Recognizing object categories observed from different viewpoints is a challenging task. Methods have to generalize along two dimensions. As illustrated on Figure 7.5 the system has to model intra-class variations (horizontal axis), as well as variations due to viewpoint changes (vertical axis) over the appearance and geometry of the given object category.

We found that current state-of-the-art techniques do not yet offer a satisfying solution for multi-view object categorization. Even the most recent methods either try to treat viewpoint changes just as a variation in appearance space, or use the "bank of detectors" approach, i.e., they train individual detectors on a discretized viewpoint space. The first approach is typically used for methods with no global spatial model, e.g., the bag-of-features representation like in (Willamowski *et al.*, 2004; Zhang *et al.*, 2007; Blaschko *et al.*, 2007). Methods using global geometry or spatial models are usually forced to learn separate models for each viewpoint.

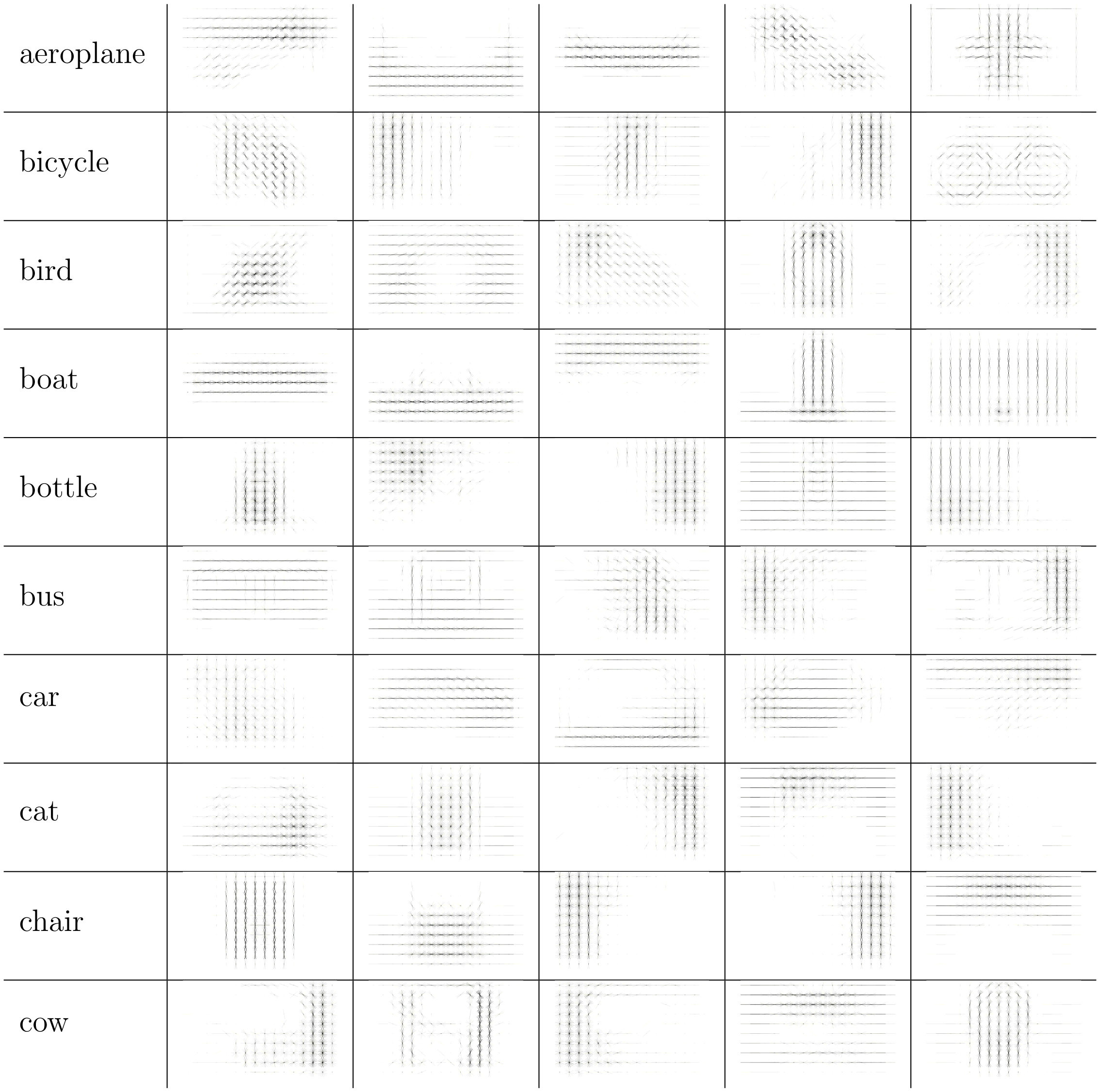


Figure 7.2: Coarse 5 topic models are estimated for each class to compute a likelihood ranking on the training images. Each row depicts the 5 topics for each class, computed on the PASCAL VOC 2007 training set. Classes aeroplane to cow

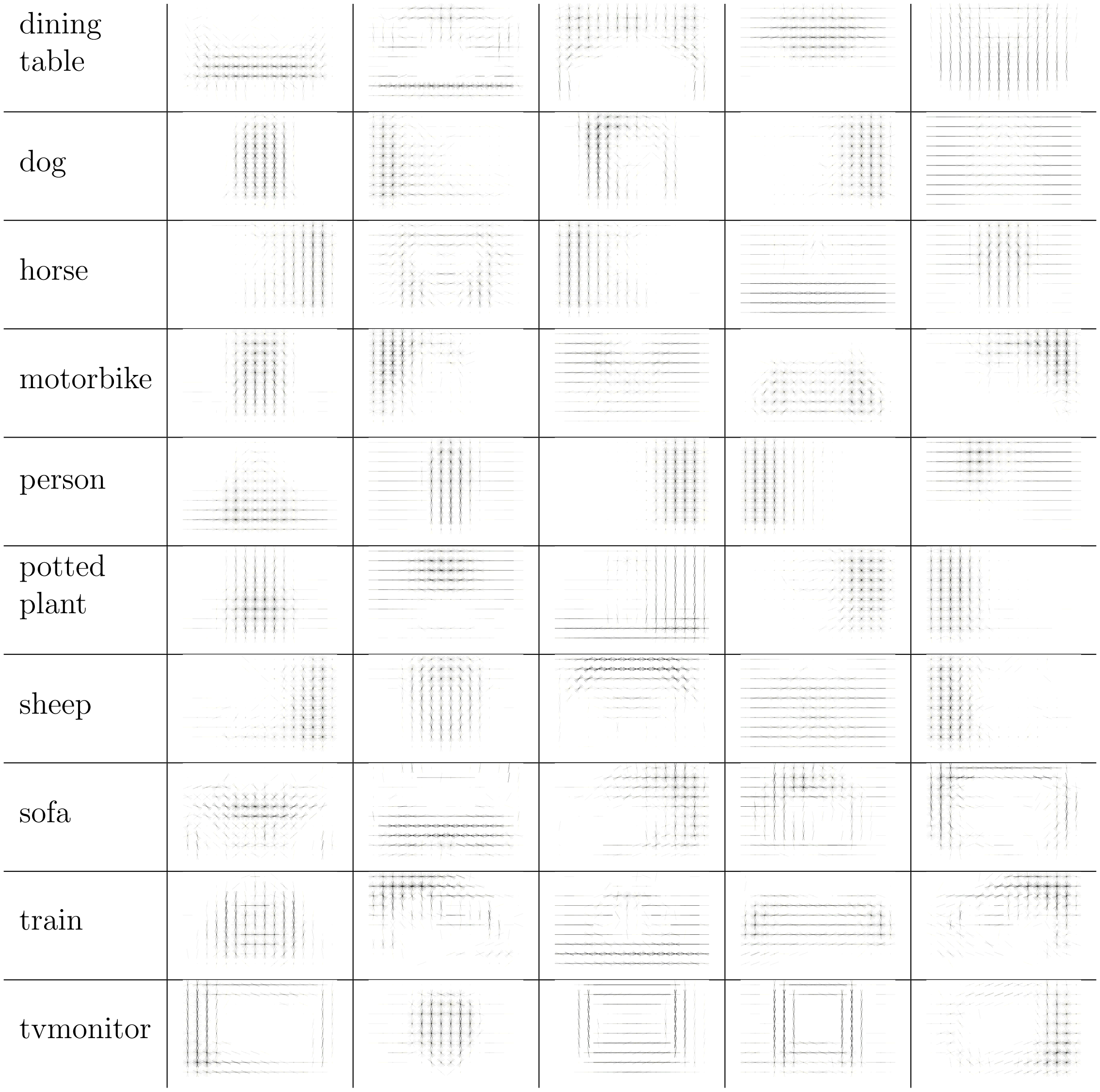


Figure 7.3: Coarse 5 topic models are estimated for each class to compute a likelihood ranking on the training images. Each row depicts the 5 topics for each class, computed on the PASCAL VOC 2007 training set. Classes dining table to tv monitor.

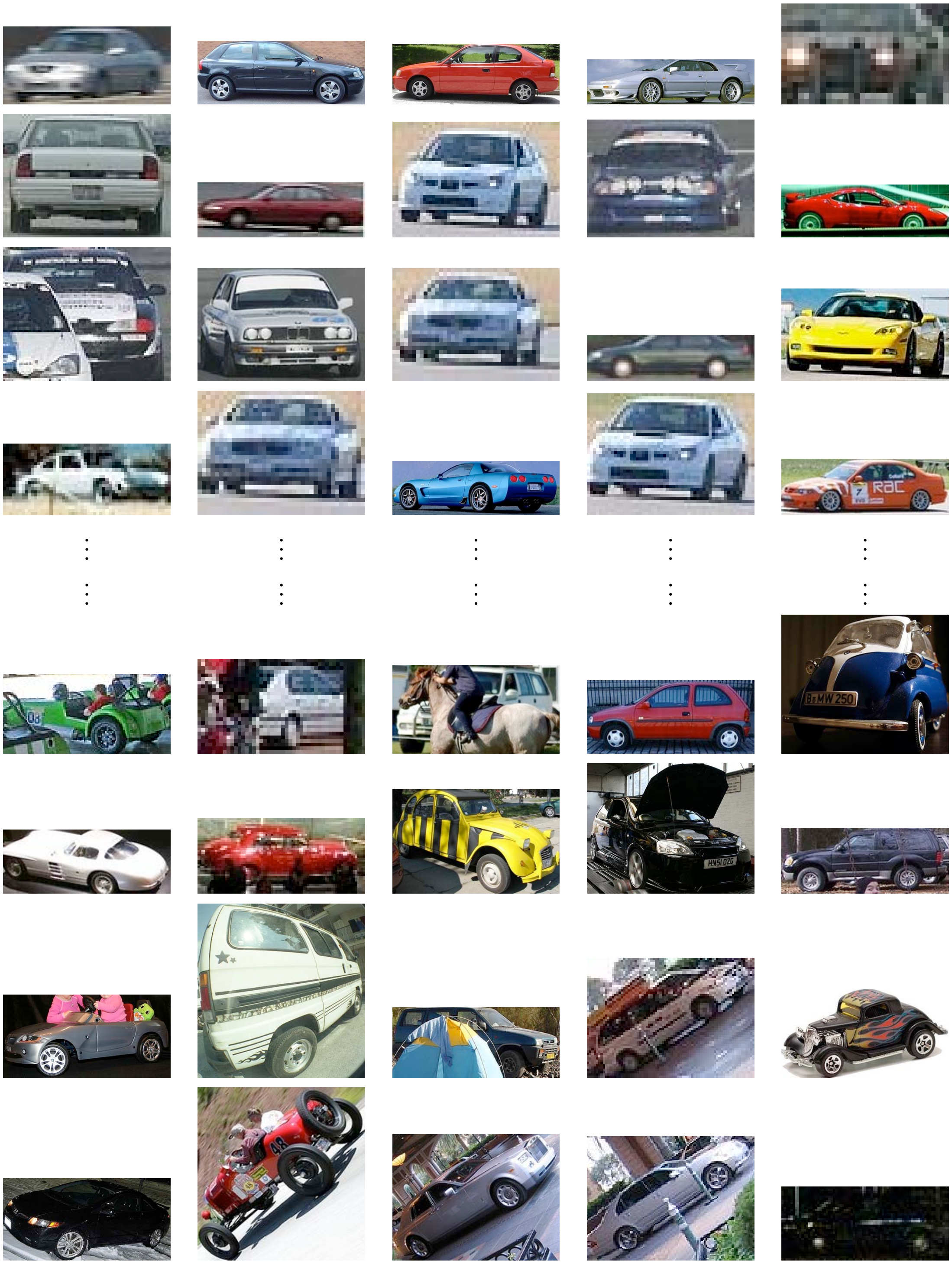


Figure 7.4: Training examples sorted according to likelihood score.

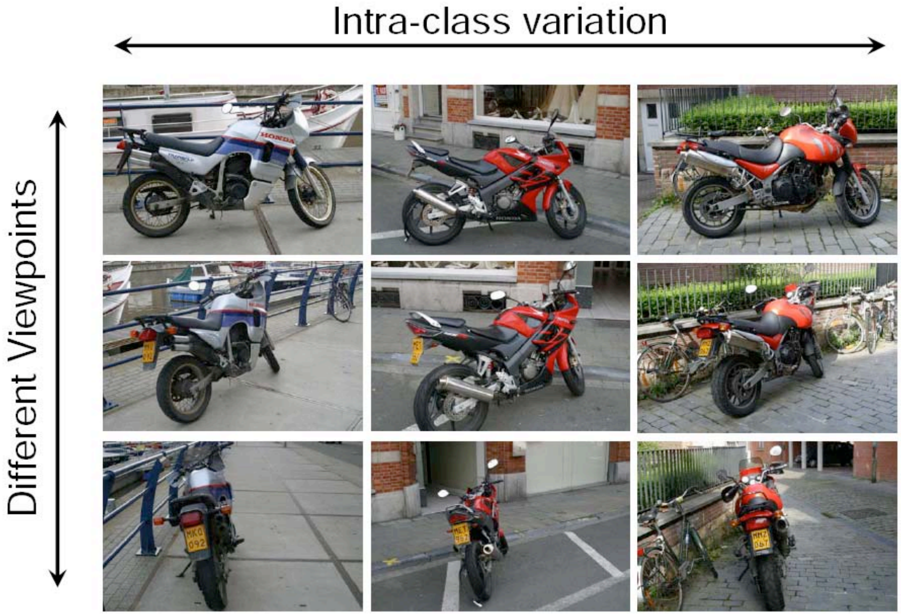


Figure 7.5: Recognizing object categories involves modeling across viewpoints and across different instances of the object categories, in this illustration motorbikes.

This typically requires efficient learning techniques and larger amount of training data. Recently, interest of the community in this topic increased and more and more object categorization method are applied to the multi-view case. Apart from a few exceptions, e.g. Thomas *et al.* (2006) who transfers (shares) object class appearance between viewpoints or Kushal *et al.* (2007), most of the methods apply robust search on possible viewpoints.

7.2.2 Unsupervised Viewpoint Discovery

Chapter 6 has concluded that topic models are a promising intermediate representation for object categories. In the following we analyze our new representation when applied to a multi-view database in more detail.

Figure 7.6 shows 20 estimated topics on multi-view point motorbike training images of Thomas *et al.* (2006). By careful inspection, one can notice that the topics automatically discovered major, including canonical, viewpoints. E.g., (k) and (p) are related to side-views, (e) and (h) are 45 degrees w.r.t side-views, and (f) are probably frontal and back views. This can be explained by that viewpoint changes effect the full descriptor, while intra-class variations can mainly be encoded by small local perturbations. The requirement of sparsity on the topic activations forces the decomposition to create topics that correspond to global patterns. In our case these global patterns correspond to object poses or viewpoints. Given these topics, a distribution over topics can be inferred for each training instance, here for each motorbike, that constitutes a decomposition. In order to further confirm our viewpoint

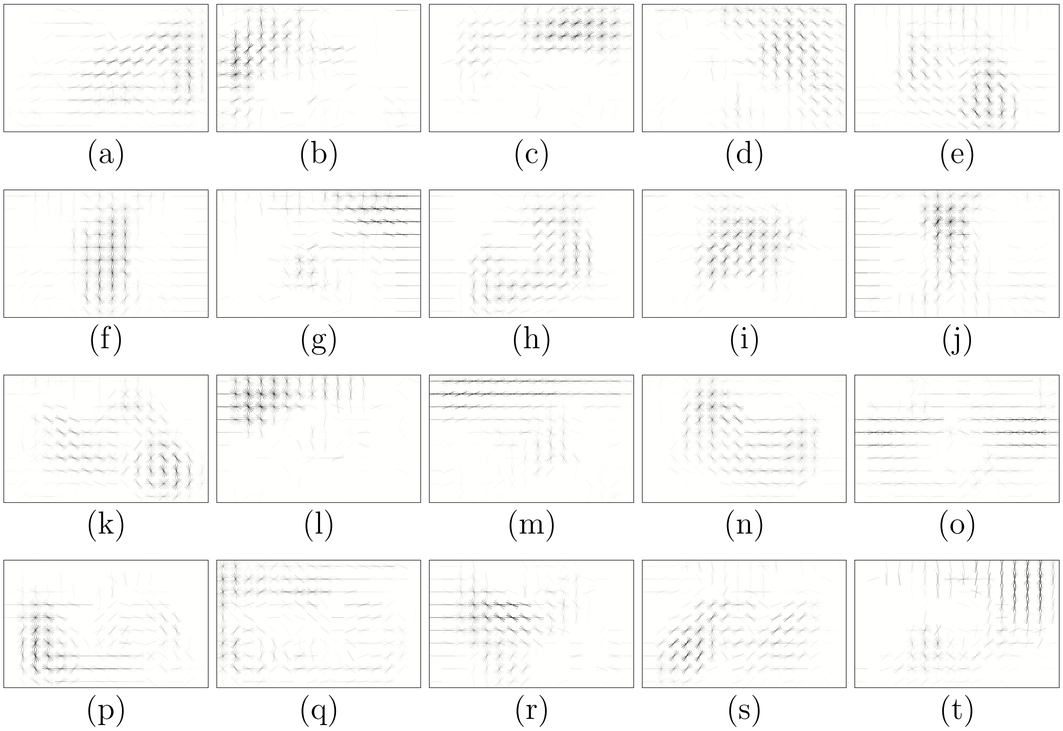


Figure 7.6: Topic decomposition on multi-view motorbikes. For explanation see the text.

discovery, Figure 7.7 (top) illustrates the space of motorbikes, where each image is represented by a vector of topic activations. We have applied multidimensional scaling to visualize the 20-dimensional latent space in two dimensions. Notice, that by moving from top to bottom, we see the viewpoint changes from frontal and back views to side-views. Also notice, that even the left and right side-views are grouped and can be well separated. To make the viewpoint transition even more apparent we have repeated the experiment with 3 topics and thus avoid possible side-effects of high-dimensional data visualization: In case of 3 topics the activation vectors lie on a simplex or on a spherical triangle depending on the normalization scheme. Results are shown on Figure 7.7 (bottom). Next section describes how this new representation can be successfully used for object recognition.

7.2.3 Bounding Box Regression

To accommodate for the varying aspect ratio due to view-point changes, we extend our model from Chapter 6 to additionally train a SVM regression model based on the inferred topic distributions to make predictions of the bounding box aspect ratio during detection. The approach is motivated by three observations. First, in Figure 6.3 we already observed, that the generative topic decomposition already discovers topics that are characteristic for certain viewpoints. For example column 8 and 12 corresponded to bicycle side-views, column 10 to car rear-views and column 13 to car

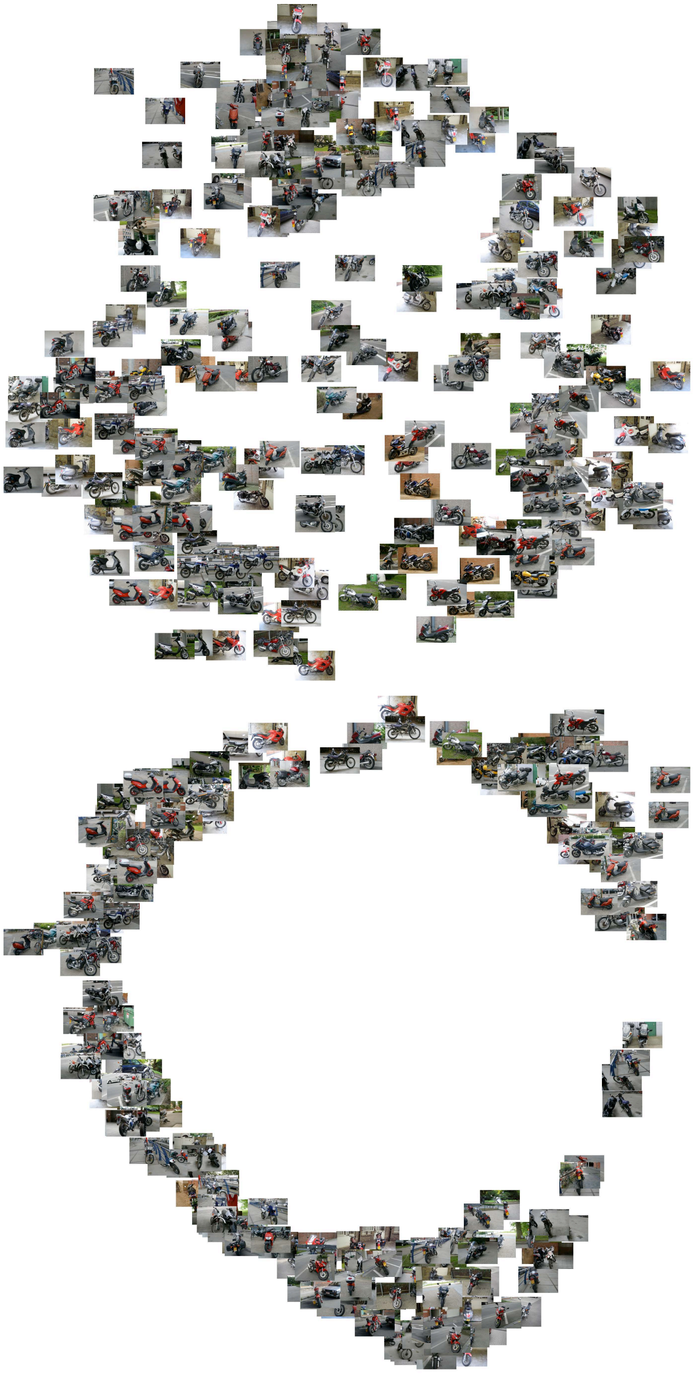


Figure 7.7: Training instances of motorbikes embedded into a 2-D space. Images are arranged based on their topic-activation-vector distance. Multidimensional scaling is used for visualization. Number of topics are 20 (top) and 3 (bottom). Instances with activation vectors around mean entropy are removed from the illustration. The layout of the morotbikes clearly demonstrate the presence topics that correspond to different viewpoints.

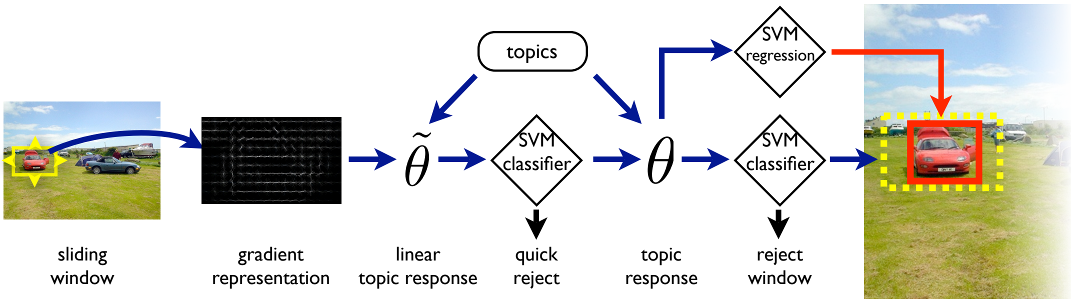


Figure 7.8: Full processing pipeline for multi-view detection using generative topic decomposition. SVM regression on the topic distribution is used to predict the aspect ratio of the unseen test example.

side-views. Second, in Section 6.2.2 we pointed out that a simple k-means clustering is able to discover distinct viewpoints of certain classes. This implies, that those examples are grouped in to some subspace by the topic representation. Third, in the last section we showed for a simpler dataset, that different viewpoints get mapped to a manifold by the topic representation that can be visualized by multidimensional scaling.

The full processing pipeline is shown in Figure 7.8. Note that the topic representation is used for detection as well as the bounding box regression. Example detections with predicted bounding box ratio are shown in Figure 7.10.

7.3 Pascal Visual Object Challenge 2007

The goal of the annual Pascal Visual Object Challenge is to provide a benchmark for recognizing objects from a number of visual object classes in realistic scenes, i.e., not pre-segmented objects. There are several visual categories in the challenge, each treated individually. We use the detection method described in 6 with the extension to data cleaning and bounding box regression from this chapter to address the car detection challenge. Our representation, similarly as in the previous section for motorbikes, is a 16x10 grid of 9 bin orientation histograms (resampled to have fixed number of words, in this case 6000). During training we estimate 200 topics for the entire training set (9963 images) using the bounding box annotation without any information of the presented categories, and a distribution over topics is inferred for each training instance as before. The detection approach, as described in Chapter 6, is based on a sliding window technique. For each hypothesized detection window the method infers a distribution over estimated topics. Based on these distributions a discriminative decision classifies the hypothesis using a support vector classifier with a chi-square kernel.

Evaluation of our method compared to all participants is shown in Figure 7.9. Our results are indicated by the “Darmstadt” curve. We can observe that our method performed fourth w.r.t. average precision, the standard evaluation criteria. This is

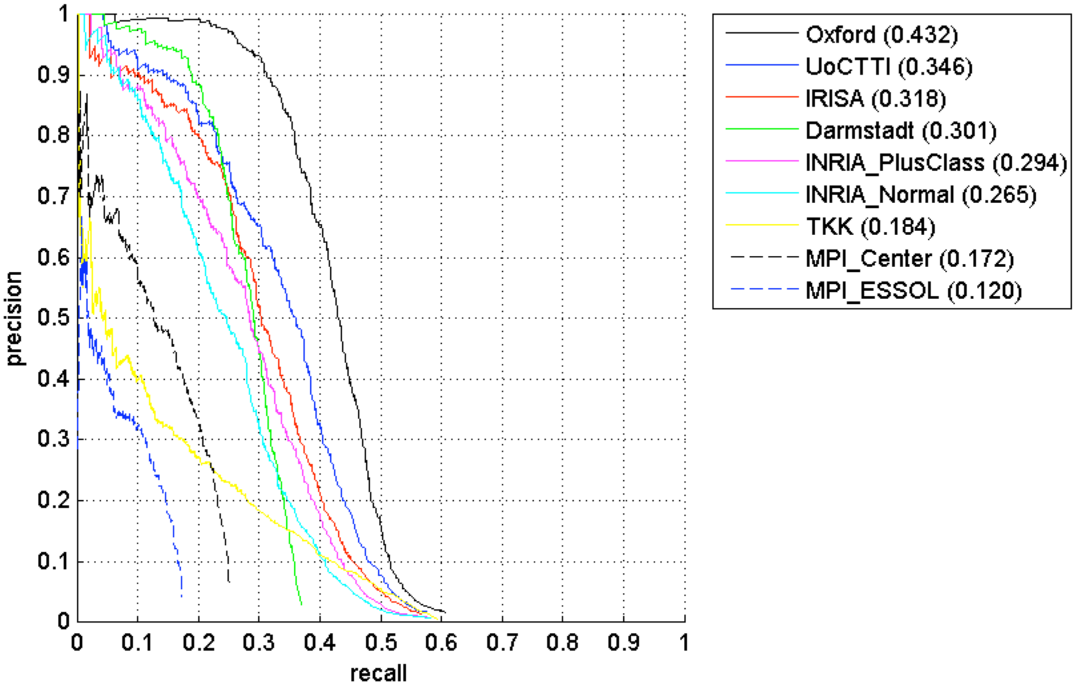


Figure 7.9: Recall-Precision curve for localization of the car class, PASCAL VOC 2007 Challenge, competition "comp3". Our results are under the "Darmstadt" label.

in our opinion considered to be a very good result on this state-of-the-art challenging dataset and by giving the strong list of competitors from the computer vision community. Moreover, by detailed inspection of the recall-precision curves, our very high precision (just below the top one) on small recall rates indicate that our most confident detection are indeed correspond to correct localization of cars. Our top 20 detections are shown in Figure 7.10. Interestingly, those best scoring detections range across different view-points, large object scale differences and some difficult illumination conditions.

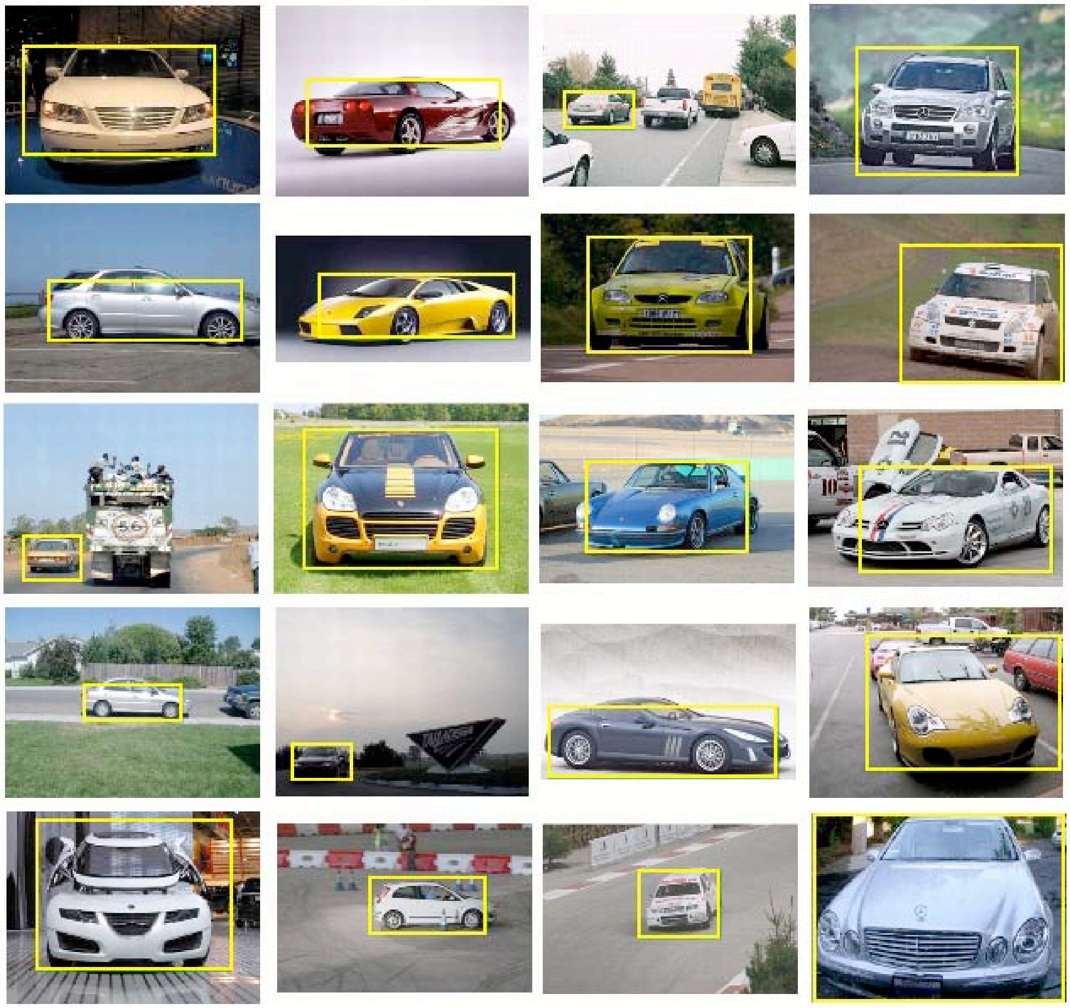


Figure 7.10: Top 20 detection of cars in the PASCAL VOC 2007.

8

Conclusion

In this thesis, we have proposed to categorize previous approaches along 3 axes which correspond to the employed modeling, representation and learning paradigms. Based on this view of the previous work, we developed methods that successfully combine different paradigms to obtain more adaptive approaches. The obtained performance improvements which we quantified by many experiments support our claims about the benefits of such adaptive methods.

8.1 Discussion of Contributions

As visual object categorization has progressed surprisingly quickly over the last few years, we considered it to be important to organize previous work in a coarse topology to make the different design choices more apparent. Thinking of different methods as employing different modeling, representation and learning paradigms became a useful tool for us to make progress towards flexible and adaptive approaches. Our scientific contributions to combine different paradigms along these 3 axes are summarized as follows:

Combining different modeling paradigms We studied the benefit of combining generative and discriminative approaches by building on the Implicit Shape Model as a generative object detector and local kernel SVMs as classifiers. We managed to combine the good generalization performance of the generative approach with the discriminative power of the discriminative classifier. The combined method therefore shows the high recall of the generative detector and as well as a reduced number of false positives and inter-category confusion due to the discriminative stage.

Combination of different learning paradigms We presented an effective coding of visual features in scale-invariant patterns that allow for efficient retrieval of reoccurring structures in a weakly supervised fashion. This scheme allows us to automatically learn annotations that can be used to train supervised detection models. The results show that we managed to obtain system performance of highly supervised methods in a weakly supervised fashion.

To also include unsupervised learning, we build on the same scale-invariant pattern representation as before which allows us to perform clustering on an object level in a data-driven manner. The obtained visual groupings of objects was employed in a semi-supervised scenario, where cross-modal learning was used to annotated the obtained clusters. Experiments show improved performance of the system by using these different levels of supervision.

Combining different representation paradigms We managed to combine different representation paradigms by learning generative decompositions of object category instances from data. The learnt components expose local, edge-based as well as global silhouette-like characteristics which are the result of learning the co-occurrence statistic via probabilistic topic models. We show the versatility and effectiveness of the obtained representation by comparing against previous approaches on unsupervised as well as supervised learning and detection tasks.

Summary We believe that this thesis contributes to the awareness of the space of design choices for visual categorization and provides a useful tool to improve methods by complementing them with aspects they miss to include. For each dimension, we provided empirical evidence that making efforts towards adaptive approaches pays off in terms of system performance instead of relying on a single spot in the design space. This relates to the intuition that there is no “sweet spot” in that space or more generally “No free lunch”. While these inherent limitations can never be fully overcome, we have shown that adaptive approaches are less brittle in this respect.

To our knowledge, the detection approach presented in Chapter 6 combines more paradigms in a single, consistent approach than previous methods we know of. The proposed representation combines local to global features by learning generative decompositions in an unsupervised way, while the detector adds a discriminative aspect by large-margin SVM classification. Of course we realize that many challenges still have to be addressed. The following section provides some perspectives, in which directions this work can be extended.

8.2 Perspectives

The main opportunities and challenges we face are related to facilitating large scale learning of far more categories. While this thesis deals with some of the key ingredients (e.g. unsupervised learning, feature sharing) which are widely believed to facilitate large scale learning, lots of computational as well as conceptual challenges remain. We will point out several aspects that follow from the presented work.

Learning from large image collections As mentioned earlier, unsupervised learning in combination with vast amounts of images as they are available today, provide a promising way towards learning far more categories. Not only does the labeling effort become increasingly prohibitive, but also labeling biases are likely to cause problems (Ponce *et al.*, 2006). To further address the inherent limitations of fully unsupervised approaches, active learning formulation (e.g. Kapoor *et al.* (2007)) seem a very promising direction to extend semi-supervised approaches as described in this thesis (see Chapter 5). But even the mere amount of data poses computational challenges. While we presented an efficient way for weakly supervised learning in Chapter 4, approaches that can handle millions of images for category learning are still missing.

Hierarchical representations A promising approach to obtain better scalability is to store information on visual categories in hierarchical (e.g. Fidler and Leonardis (2007)) or at least a sparse graph-based structures. It is yet unclear how to construct and exploit those structures for maximum accuracy, storage efficiency or speed. Recently, extensions of the probabilistic model we used in Chapter 6 (Blei *et al.*, 2003a) have been used to learn taxonomies from data (Sivic *et al.*, 2008; Bart *et al.*, 2008). Leveraging the recent success of hierarchy learning, our approach can be extend to learn hierarchical decompositions, opening up new opportunities for compact representations.

Deformation Modeling One of the draw-backs we addressed in Chapter 6 is the rigid reference frame, which tends to work better on rigid object categories in comparison to articulated and deformable object categories. The presented model can be augmented by a deformation model that accounts for variation in the position of the sub-structures. Preliminary experiments have shown that balancing deformation invariance vs. discriminance is a non-trivial problem (see also Varma and Ray (2007) for a related discussion). Nevertheless, such developments could lead to less redundant representations that also capture the dynamics of articulated categories.

List of Figures

1.1 Example of high intra-class variation between the dogs and low inter-class variation to instances of other categories. 2

2.1 The two key ingredients of the Implicit Shape Model: (a) visual codebook vocabulary and (b) spatial occurrence distribution of the codebook entries. Both statistics are computed on the Caltech motorbike dataset. 16

2.2 Illustration of the normal form of a hyperplane. The orientation of the hyperplane is given by the orthogonal vector $\vec{w}$ which is constraint to $||\vec{w}|| = 1$ and the distance from the origin by b. The hyperplane is chosen to separate the two shown classes with a maximal margin. The so-called support vectors are marked in grey. 19

2.3 Noisy data is handled by slack variables ξ_i which allows some data points to lie within the margin or even on the "wrong" side of the hyperplane. 21

2.4 Illustration of a non-linear mapping Φ which makes the classes in the data linearly separable. 22

2.5 LDA model as formulated by Griffiths and Steyvers (2004). 23

3.1 Stages of the integrated approach. (a) original image; (b) hypotheses detected by the representative ISM; (c) input to the SVM stage; (d) verified hypotheses. 29

3.2 A look inside the LK verification stage. Each support vector specifies a configuration of local features, corresponding to a particular training example. When evaluating a hypothesis, the kernel K first searches for the k best feature correspondences, considering both appearance and relative position, and uses them to judge the quality of the match. 32

3.3 Recall-Precision curves for the car, cow, horse and motorbike model on a detection task. 35

3.4 Cross-category confusions (false positives per test image) for the ISM and our new IRD approach on a detection task. 36

3.5 Example detections on the car, cow, horse, and motorbike test sets. . 36

3.6 Precision-Recall curves for the difficult multi-scale databases of cars and motorbikes. Our new IRD approach clearly outperforms the state-of-the-art on the UIUC multi-scale car database. 37
3.7 Example detections of our new IRD approach on the difficult multi-scale UIUC car database and the multi-scale TUD motorbike test set. 38
3.8 Results on test sets 1 (left) and 2 (right) for motorbikes.The top row shows the performance for the object present/absent classification task. The bottom row contains our results for the object detection task. 40
3.9 Results on test sets 1 (left) and 2 (right) for cars.The top row shows the performance for the object present/absent classification task. The bottom row contains our results for the object detection task. 40
3.10 Results on the PASCAL Challenge motorbike detection task. 41
3.11 Example detections and segmentations on PASCAL Challenge motorbike detection task. 41
3.12 Example detections and segmentations on PASCAL Challenge car detection task. 42
3.13 Results on the PASCAL Challenge car detection task. 42

4.1 Cluster centers of the generic codebook. The codebook is unspecific since no knowledge about the particular object class is assumed during object discovery. 50 codebook entries are obtained by k-means clustering on the Caltech background dataset using a Hessian-Laplace interest point detector and gray value patches as descriptors. 46
4.2 The scale-invariant representation we use is derived from the common bag-of-words representation. Recently a spatial resolution was added to each bin. We improve the robustness of these type of representations by adding a scale-normalization step yet retaining the spatial information which yields what we call a Scale-Invariant Pattern. . . . 47
4.3 Example of the computed likelihood on the multi-scale TUD motorbikes. 49
4.4 Example result of our procedure for object discovery on car and cow images including varying position, scale and viewpoint and heterogeneous background. 50
4.5 Quantitative evaluation of the error in the center point estimation on the Caltech motorbike training set. (a) shows the distribution of the pixel distance between annotation and estimate and (b) shows the distribution of the distance normalized by the object width. 51
4.6 The dependencies between the unknown object scales s and the estimated relative scales ρ is visualized as a fully connected graph. After pruning of the graph, estimation of the object scales s are estimated by least squares. 52

4.7 Evaluation of the scale estimate on the Caltech database. The 400 images are sorted by the true scale of the object. The scale estimate for each of the proposed methods is plotted logarithmically. The curve for the pure least-square and minimum linkage method are virtually identical. 54

4.8 The proposed method for Object Discovery also facilitates ranking of the images. (top) best ranked images (bottom) worst ranked images. 57

4.9 (a) training image (b) estimated approximation of object segmentation (c) test image (d)+(e) inferred segmentation for hypothesis (f) final detections 58

4.10 (left) ROC-curve of ranking task (middle) performance comparison to supervised baseline and performance increase due to approximated segmentations (right) generalization to other categories and data sets 59

5.1 System overview. 63

5.2 Multi-dimensional scaling using scale-invariant pattern representation. (left) 4 of the 5 classes (mobile, pen, bottle, can, apple) are clearly separated without using any supervision and despite the small statistic. Only the categories bottle and can are confused. (right) all 4 caltech classes (airplane, face, leave, motorbike) are well separated using 200 examples from each category. The groundtruth labels are displayed color-coded. 65

5.3 Visual grouping of objects by clustering. 66

5.4 One basic approach to semi-supervised learning. Label information is propagated from labeled samples to unlabeled ones based on similarity of the observed samples. Note the refined decision boundary implied by this process. 68

5.5 4 non-parametric probability density functions for modeling the spatial relations $p(\text{pos}(\Psi_i), \text{pos}(\Psi_j)|R)$ between feature patterns in 2d image coordinates, where R corresponds to one of the spatial relations: (a) left of (b) right of (c) above (d) below 71

5.6 Semi-supervised learning is implemented in our system by assigning labels to the clusters only with a certain belief that represents the uncertainty resulting from the dialogue. Hereby, the system can recover from erroneous beliefs. The columns in the presented probability table represented the categories communicated via the language system and the rows correspond to the clusters obtained by the visual grouping step. The cluster members are visualized on the right. 72

5.7 (a) shows an example scenario for propagation of labels from known categories to unknown ones. (b) shows the improvement of the semi-supervised approach over the purely supervised approach by exploiting information from unlabeled data. 73

5.8 Scenario which shows how the system updates associations (light green) to recover form an incorrect belief. 74

6.1 Dense gradient histogram representation. The loupe shows the 9 possible edge orientation of the histogram bins that are interpreted as words. 78

6.2 First row: example topics of 8 topic model for classes airplane, face, motorbike, watch. Second row: example topics of 50 topic model for the same classes. 79

6.3 First row: example topics that were learned by the proposed approach across categories and viewpoints for the 10 classes of the PASCAL'06 data. Below first row: training images that activated the topic above most. The topics model local structures, line segments as well as silhouette-like structures. The topics are distinctive enough to separate several category members and even view-points. On the other hand they are general enough to be shared across categories and viewpoints. 82

6.4 Unsupervised discovery of categories and viewpoints in PASCAL'06 data. The rows show for all 50 clusters those 10 examples that are closest to the cluster center. The left block visualizes the clusters 1 to 25 and the right block visualizes the clusters 26-50. 83

6.5 (a) and (b): Comparison of learning curve for proposed intermediate representation versus SVM on pure features on UIUC single-scale database. 86

6.6 (a) Performance on UIUC multi-scale dataset using topic model estimated via Gibbs sampling vs variational bayes approach compared to using pseudo topic activations. (b)Precision-Recall curves on the PASCAL VOC challenge 2006. Precision-Recall curves and example detections. 88

6.7 Example detections on the PASCAL VOC challenge 2006. 88

6.8 Example topics of 100 topic model jointly learned on apple-logos, bottles, giraffes, mugs and swans. 89

6.9 Example detections on the ETH shape database. 90

7.1 PASCAL'07 training set contains examples that expose heavy occlusion and lighting artefaces. 93

7.2 Coarse 5 topic models are estimated for each class to compute a likelihood ranking on the training images. Each row depicts the 5 topics for each class, computed on the PASCAL VOC 2007 training set. Classes aeroplane to cow 94

7.3 Coarse 5 topic models are estimated for each class to compute a likelihood ranking on the training images. Each row depicts the 5 topics for each class, computed on the PASCAL VOC 2007 training set. Classes dining table to tv monitor. 95

7.4 Training examples sorted according to likelihood score. 96

7.5 Recognizing object categories involves modeling across viewpoints and across different instances of the object categories, in this illustration motorbikes. 97

7.6 Topic decomposition on multi-view motorbikes. For explanation see the text. 98

7.7 Training instances of motorbikes embedded into a 2-D space. Images are arranged based on their topic-activation-vector distance. Multidimensional scaling is used for visualization. Number of topics are 20 (top) and 3 (bottom). Instances with activation vectors around mean entropy are removed from the illustration. The layout of the morotbikes clearly demonstrate the presence topics that correspond to different viewpoints. 99

7.8 Full processing pipeline for multi-view detection using generative topic decomposition. SVM regression on the topic distribution is used to predict the aspect ratio of the unseen test example. 100

7.9 Recall-Precision curve for localization of the car class, PASCAL VOC 2007 Challenge, competition "comp3". Our results are under the "Darmstadt" label. 101

7.10 Top 20 detection of cars in the PASCAL VOC 2007. 102

List of Tables

3.1 Equal error rate performances achieved by the Local Kernel (LK), Implicit Shape Model (ISM), and our integrated approach (IRD) on present/absent tasks. 34

3.2 Analysis of the image sources for the car test set 1. Only the TUD and UIUC databases contain exclusively side views. Altogether, side views take up only about 55% of the test cases. 39

4.1 Noise analysis of our new object scale estimation technique. Different methods to obtain the final estimate are evaluated at different noise levels of the annotation of the object center. The error in the scale estimate is given in mean squared deviation form the groundtruth in log-space. 54

6.1 Comparison to other approaches on re-ranking task of google images. Performance is measured in precision at 15% recall. In contrast to the other methods our approach does not use any validation set. 80

6.2 Average precision achieved on the PASCAL'06 database. 87

6.3 Comparison against shape-based approach of Ferrari *et al.* (2007) on ETH shape database. Average detection-rate at 0.4 false positives per image averaged over 5-folds. Standard deviation is specified in brackets. 89

Bibliography

A. Agarwal and B. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In , *European Conference on Computer Vision (ECCV'06)*, volume 3951 of *Lecture Notes in Computer Science*, Graz, Austria, May 2006. Springer. 64

S. Agarwal, A. Atwan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004. 28, 29, 34, 35, 37

A. Arsenio. Developmental learning on a humanoid robot. In *International Joint Conference On Neural Networks*, 2004. 63

J. Baldridge and G.J.M. Kruijff. Coupling ccg and hybrid logic dependency semantics. In *ACL '02*, Morristown, NJ, USA, 2001. 69

J. Baldridge and G.J.M. Kruijff. Multi-modal combinatory categorial grammar. In *EACL '03*, Morristown, NJ, USA, 2003. 69

E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005. IEEE Computer Society. 9, 63

E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, Anchorage, Alaska, USA, June 2008. IEEE Computer Society. 105

C. Bauckhage, G.A. Fink, J. Fritsch, F. Kummert, F. Lomker, and G.S. andS. Wachsmuth. An integrated system for cooperative man-machine interaction. In *Computational Intelligence in Robotics and Automation*, 2001. 62

S. Belongie, J. Malik, and J. Puchiza. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(4), April 2002. 16

I. Biederman. Recognition by components: A theory of human image understanding. *Psychol. Review*, 94:115–147, 1987. 1

C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007. 9, 10

A. Bissacco, M.H. Yang, and S. Soatto. Detecting humans via their pose. In *Advances in Neural Information Processing Systems (NIPS'07)*. MIT Press, Cambridge, MA, 2008. 22, 25

M.B. Blaschko, T. Hofmann, and C.H. Lampert. Efficient subwindow search for object localization. Technical Report 164, Max Planck Institute for Biological Cybernetics, August 2007. 93

D. Blei, T. Gri, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Neural Information Processing Systems(NIPS)*, 2003. 105

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003. 15, 22, 23, 84, 86

E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In , *European Conference on Computer Vision (ECCV'02)*, volume 2350 of *Lecture Notes in Computer Science*, pages 109–122, Copenhagen, Denmark, May 2002. Springer. 34

A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In , *European Conference on Computer Vision (ECCV'06)*, volume 3951 of *Lecture Notes in Computer Science*, Graz, Austria, May 2006. Springer. 22

G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 710–715, San Diego, CA, USA, June 2005. IEEE Computer Society. 10

S. Boughorbel, J.P. Tarel, and F. Fleuret. Non-mercer kernels for svm object recognition. In *British Machine Vision Conference (BMVC'04)*, pages 137 – 146, London, UK, September 2004. British Machine Vision Association. 28

M.C. Burl and P. Perona. Recognition of planar object classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96)*, page 223, San Francisco, CA, USA, June 1996. IEEE Computer Society. 7, 8, 13

L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 15

G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing Systems (NIPS'00)*, pages 409–415, Cambridge, MA, 2001. MIT Press. 10

C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 81

O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, May 1999. 21

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006. 15

H.P. Chiu, L.P. Kaelbling, and T. Lozano-Pérez. Virtual training for multi-view object class recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 12

O. Chum and A. Zisserman. An exemplar model for learning object classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 14, 87

D. Comaniciu and P. Meer. Distribution free decomposition of multivariate data. *Pattern Analysis and Applications*, 2(1):22–30, 1999. 17

D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *IEEE International Conference on Computer Vision (ICCV'01)*, Vancouver, Canada, July 2001. IEEE Computer Society. 17

T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In , *European Conference on Computer Vision (ECCV'98)*, volume 1406 of *Lecture Notes in Computer Science*, Freiburg, Germany, June 1998. Springer. 7

N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines: and other kernel-based learning methods.* Cambridge University Press, 2000. 21

G. Csurka, C. Dance, L. Fan, J. Willarnowski, and C. Bray. Visual categorization with bags of keypoints. In *SLCV'04*, pages 59–74, Prague, Czech Republic, May 2004. 13, 47, 64

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005. IEEE Computer Society. 8, 9, 13, 14, 25, 43, 78, 84, 87

D. Decoste and B. Schölkopf. Training invariant support vector machines. *Mach. Learn.*, 46(1-3):161–190, 2002. 9, 12

G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *IEEE International Conference on Computer Vision (ICCV'03)*, Nice, France, October 2003. IEEE Computer Society. 10, 12

M. Everingham and A. Zisserman. Identifying individuals in video by combining "generative" and discriminative head models. In *IEEE International Conference on Computer Vision (ICCV'05)*, pages 1103–1110, Beijing, China, October 2005. IEEE Computer Society. 12

M. Everingham, A. Zisserman, C.K.I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2005 (VOC2005) Results. http://www.pascal-network.org/challenges/VOC/voc2005/results.pdf, 2005. 34, 48, 49

M. Everingham, A. Zisserman, C.K.I. Williams, L.J.V. Gool, M. Allan, C.M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J.D.R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A.J. Storkey, S. Szedmák, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 pascal visual object classes challenge. In *First PASCAL Machine Learning Challenges Workshop MLCW*, pages 117–176, 2005. 33

M. Everingham, A. Zisserman, C.K.I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006. 80, 87

M. Everingham, A. Zisserman, C.K.I. Williams, J. Winn, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007, 2007. 1, 91

L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 1134–1141, Nice, France, October 2003. IEEE Computer Society. 9

L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1134, Washington, DC, USA, 2003. IEEE Computer Society. 63

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, volume 12, page 178, Los Alamitos, CA, USA, 2004. IEEE Computer Society. 10

P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55–79, 2005. 9, 45

R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR'03), Madison, WI, USA, June 2003. IEEE Computer Society. 1, 7, 8, 9, 12, 13, 14, 24, 28, 34, 45, 46, 49, 54, 59, 61

R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 14, 15, 22, 56, 79, 80

R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision (IJCV)*, 2005. 59

V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 89, 113

V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2008. to appear. 13

S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 10, 105

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Europ. Conf. on Comput. Learning Theory*, pages 23–37, 1995. 8

J.H. Friedman. On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, 1(1):55–77, 1997. 10, 11

M. Fritz and B. Schiele. Towards unsupervised discovery of visual categories. In *DAGM Symposium Symposium for Pattern Recognition (DAGM'06)*, Berlin, Germany, September 2006. 3, 4, 64

M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, Anchorage, Alaska, USA, June 2008. IEEE Computer Society. 3, 5

M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 2, 4, 59, 77, 87

M. Fritz, G.J.M. Kruijff, and B. Schiele. Cross-modal learning of visual categories using different levels of supervision. In *Proceedings of 5th International Conference on Computer Vision Systems*, 2007. 3, 5

D.M. Gavrila and L.S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96)*, page 73, Washington, DC, USA, June 1996. IEEE Computer Society. 7

D. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *International Conference on Pattern Recognition (ICPR'98)*, volume 1, pages 439–444, 1998. 12, 13

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. CRC Press, 2004. 10

K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006. IEEE Computer Society. 61, 66

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 1

T.L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS USA*, 2004. 22, 23, 84, 86, 107

S. Harnad. The symbol grounding problem. *Phys. D*, 42(1-3):335–346, 1990. 62

N. Hawes, M. Brenner, and K. Sjöö. Planning as an architectural control mechanism. In *Proceedings of Human-Robot Interaction 2009 (HRI '09)*, March 2009. 70

E. Hayman, B. Caputo, M. Fritz, and J.O. Eklundh. On the significance of real-world conditions for material classification. In , *European Conference on Computer Vision (ECCV'04)*, volume 3021 of *Lecture Notes in Computer Science*, Prague, Czech Republic, May 2004. Springer. 91

B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 657–662, Kauai, HI, USA, December 2001. IEEE Computer Society. 28, 29

A.B. Hillel, T. Hertz, and D. Weinshall. Efficient learning of relational object class models. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 11, 49

T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001. 15, 22

A.D. Holub, M. Wellling, and P. Perona. Combining generative models and fisher kernels for object recognition. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 12

A.D. Holub, M. Welling, and P. Perona. Hybrid generative-discriminative visual categorization. *International Journal of Computer Vision (IJCV)*, 77(1-3):239–258, 2008. 9, 15

T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press. 10, 11, 24, 27, 77

H. Jacobsson, N. Hawes, G. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008. 69

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004. 11, 27

T. Jebara. *Discriminative, generative and imitative learning.* PhD thesis, MIT, 2002. Supervisor-Alex P. Pentland. 8

A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 12, 15, 105

J. Kelleher, G. Kruijff, and F. Costello. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of ACL/COLING 2006*, 2006. 70

J. Kelleher, G.J. Kruijff, and F. Costello. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expression. In *Coling-ACL '06*, Sydney Australia, 2006. 71

S. Kirstein, H. Wersing, and E. Körner. Rapid online learning of objects in a biologically motivated recognition architecture. In *27th Pattern Recognition Symposium DAGM*, pages 301–308. Springer, 2005. 63

G.J.M. Kruijff, J.D. Kelleher, G. Berginc, and A. Leonardis. Structural descriptions in Human-Assisted robot visual learning. In *Proceedings of 1st Annual Conference on Human-Robot Interaction*, March 2006. 62, 69

G.J.M. Kruijff, J.D. Kelleher, and N. Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In *PIT 2006*, Kloster Irsee, Germany, June 2006. 69

G.J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes. Situated dialogue processing for human-robot interaction. In , *Cognitive Systems*. Springer Verlag, 2009. 70

A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 97

G. Lakoff. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. The University of Chicago Press, Chicago, 1987. 7

I. Laptev. Improvements of object detection using boosted histograms. In *British Machine Vision Conference (BMVC'06)*, Edinburgh, UK, September 2006. British Machine Vision Association. 13

D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *British Machine Vision Conference (BMVC'06)*, Edinburgh, UK, September 2006. British Machine Vision Association. 22

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178, New York, NY, USA, June 2006. IEEE Computer Society. 47, 64

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 7

B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *SLCV'04*, pages 17–32, Prague, Czech Republic, May 2004. 8, 9, 12, 17, 27, 28, 30, 45, 55, 59

B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005. IEEE Computer Society. 1, 7, 12, 61, 64, 77

B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision (IJCV)*, 77(1-3):259–289, 2008. 13, 14, 15, 17, 24, 28, 46, 55

K. Levi and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, June 2004. IEEE Computer Society. 13

F.F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005. IEEE Computer Society. 22

Y. Li, L.G. Shapiro, and J.A. Bilmes. A generative/discriminative learning algorithm for image classification. In *IEEE International Conference on Computer Vision (ICCV'05)*, pages 1605–1612, Beijing, China, October 2005. IEEE Computer Society. 11

L.J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 15

T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision (IJCV)*, 30(2):79–116, 1998. 13

P. Lison and G. Kruijff. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of ECAI 2008*, Athens, Greece, 2008. 69

P. Lison and G.J.M. Kruijff. An integrated approach to robust processing of situated spoken dialogue. In *Proceedings of the Second International Workshop on the Semantic Representation of Spoken Language (SRSL'09)*, Athens, Greece, 2009. 69

D. Lowe. Object recognition from local scale invariant features. In *IEEE International Conference on Computer Vision (ICCV'99)*, Kerkyra, Greece, September 1999. IEEE Computer Society. 64

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 12, 15, 16, 28, 30, 55

J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. Incremental learning for place recognition in dynamic environments. In *Intelligent Robots and Systems (IROS'07)*, pages 721–728, San Diego, CA, USA, 2007. 10

D.R. Martin, C.C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. 13

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03)*, pages II: 257–263, Madison, WI, USA, June 2003. IEEE Computer Society. 64

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 13, 47, 55

K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 64

K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006. IEEE Computer Society. 12, 13, 87

A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. Pattern Anal. Mach. Intell*, 23(4):349–361, 2001. 13

P.J. Moreno, P.P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In , *Advances in Neural Information Processing Systems (NIPS'04)*. MIT Press, Cambridge, MA, 2005. 11

H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision (IJCV)*, 14:5–24, 1995. 7

J. Mutch and D.G. Lowe. Multiclass object recognition with sparse, localized features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006. IEEE Computer Society. 87

A.Y. Ng and M.I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems (NIPS'01)*, pages 841–848, Cambridge, MA, 2002. MIT Press. 9, 10, 11

M. Nilsback and B. Caputo. Cue integration through discriminative accumulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, June 2004. IEEE Computer Society. 8, 10, 27, 28, 29

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In , *European Conference on Computer Vision (ECCV'06)*, volume 3951 of *Lecture Notes in Computer Science*, Graz, Austria, May 2006. Springer. 14

S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. Bakir. Weighted substructure mining for image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 14

A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 3–10, New York, NY, USA, June 2006. IEEE Computer Society. 10

E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, San Juan, Puerto Rico, June 1997. IEEE Computer Society. 87

C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision (IJCV)*, 38(1):15–33, 2000. 28, 29

J. Ponce, T.L. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, C. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Towards Category-Level Object Recognition*, pages 29–48. Springer, 2006. 14, 91, 105

M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(6):637–646, 1998. 28

T. Quack, V. Ferrari, B. Leibe, and L.V. Gool. Efficient mining of frequent and distinctive feature configurations. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 14

P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L.V. Gool. Modeling scenes with local descriptors and latent aspects. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 22, 25

E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976. 7

D. Roy. Learning words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002. 62, 69

B.C. Russell, W.T. Freeman, A.A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1605–1614, New York, NY, USA, June 2006. IEEE Computer Society. 15

B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, January 2000. 7

H. Schneiderman and T. Kanade. A statistical method of 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, pages 746–751, Hilton Head, SC, USA, June 2000. IEEE Computer Society. 7, 29

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001. 8, 15, 18, 20

F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 15, 80

E. Seemann and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *British Machine Vision Conference (BMVC'05)*, Oxford, UK, September 2005. British Machine Vision Association. 16

J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their locations in images. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 14, 15, 22, 24, 25, 61, 64, 65, 79

J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, and A.A. Efros. Unsupervised discovery of visual object class hierarchies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, Anchorage, Alaska, USA, June 2008. IEEE Computer Society. 105

D. Skočaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 1494–1501, Nice, France, October 2003. IEEE Computer Society. 10

D. Skočaj, G. Berginc, B. Ridge, A. Štimec, M. Jogan, O. Vanek, A. Leonardis, M. Hutter, and N. Hewes. A system for continuous learning of visual concepts. In *International Conference on Computer Vision Systems ICVS 2007*, Bielefeld, Germany, March 2007. 62, 63

G. Socher, G. Sagerer, and P. Perona. Bayesian reasoning on qualitative descriptions from images and speech. In *Image and Vision Computing*, 2000. 62

L. Steels and F. Kaplan. Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2001. 63

L. Steels. The symbol grounding problem has been solved. so what's next? In , *Symbols, embodiment and meaning*. Academic Press, New Haven, 2008. 62, 69

M. Steyvers and T.L. Griffiths. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007. 23, 85

E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *IEEE International Conference on Computer Vision (ICCV'05)*, Beijing, China, October 2005. IEEE Computer Society. 10, 22

A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L.V. Gool. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006. IEEE Computer Society. 39, 97

A.L. Thomaz. *Socially Guided Machine Learning*. PhD thesis, Massachusetts Institute of Technology, May 2006. 68, 69

A. Torralba, K. Murphy, and W. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Washington, DC, USA, June 2004. IEEE Computer Society. 8, 10, 29

A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(5), 2007. 25, 77, 81

K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.R. Müller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10):2397–2414, 2002. 27

Z. Tu, X. Chen, A.L. Yuille, and S.C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 18–25, Nice, France, October 2003. IEEE Computer Society. 12

M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991. 7

I. Ulusoy and C.M. Bishop. Generative versus discriminative methods for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, June 2005. IEEE Computer Society. 10

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1996. 10, 18, 20

M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 1, 7, 9, 105

N. Vasconcelos, P. Ho, and P.J. Moreno. The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition. In , *European Conference on Computer Vision (ECCV'04)*, volume 3021 of *Lecture Notes in Computer Science*, pages 430–441, Prague, Czech Republic, May 2004. Springer. 12, 27

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 511–518, Kauai, HI, USA, December 2001. IEEE Computer Society. 7, 14

P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004. 8, 28, 29

P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *IEEE International Conference on Computer Vision (ICCV'03)*, pages 734–741, Nice, France, October 2003. IEEE Computer Society. 10, 28

C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *IEEE International Conference on Computer Vision (ICCV'03)*, Nice, France, October 2003. IEEE Computer Society. 24, 27, 28, 29, 32

M. Weber, M. Welling, and P. Perona. Unsupervised learning of object models for recognition. In , *European Conference on Computer Vision (ECCV'00)*, volume 1843 of *Lecture Notes in Computer Science*, Dublin, Ireland, June 2000. Springer. 14, 45

J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *International Workshop on Learning for Adaptable Visual Systems (LAVS04), Cambridge, United Kingdom*, August 2004. 93

J.M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *IEEE International Conference on Computer Vision (ICCV'05)*, pages 756–763, Beijing, China, October 2005. IEEE Computer Society. 14

H.J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Comput. Sci. Eng.*, 4(4):10–21, 1997. 48, 66

B. Wu and R. Nevatia. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *IEEE International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil, October 2007. IEEE Computer Society. 87

Z. Yao, X. Yang, and S. Zhu. Introduction to a large scale general purpose groundtruth dataset: methodology, annotation tool, and benchmarks. In *International Conference on Energy Minimization Methods in CVPR (EMMCVPR)*, 2007. 91

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision (IJCV)*, 73(2):213–238, 2007. 93

X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf. 15

Curriculum Vitae

Mario Fritz

Date of birth:	1/16/1978 in Adenau, Germany
Citizenship:	German

Positions:

2004–2008 **TU-Darmstadt, Germany**

PhD student, supervisor: Prof. Bernt Schiele

2003–2004 **KTH, Stockholm, Sweden**

Visiting researcher, host: Prof. Jan-Olof Eklundh

Education:

2001–2008 **Distance University of Hagen, Germany**

Study of business administration

Majors: statistics, operations research

1998–2004 **University of Erlangen-Nürnberg, Germany**

Study of computer science, degree: Diplom-Informatiker

Majors: pattern recognition, scientific simulation, computer graphics, math

Master thesis: ”Categorization by Local Information using Support Vector Machines” supervised by Barbara Caputo

Student thesis: ”3D Object Tracking using Light-Fields” supervised by Matthias Zobel

---

Publications

[11] Paul Schnitzspan, Mario Fritz, and Bernt Schiele.
Hierarchical support vector random fields: Joint training to combine local and global features.
In European Conference on Computer Vision **ECCV'08**, 2008. to appear.

[10] Tâm Huỳnh, Mario Fritz, and Bernt Schiele.
Discovery of activity patterns using topic models.
In International Conference on Ubiquitous Computing **UBICOMP'08**, 2008. to appear.

[9] Mario Fritz and Bernt Schiele.
Decomposition, discovery and detection of visual categories using topic models.
In IEEE Conference on Computer Vision and Pattern Recognition **CVPR'08**, June 2008.

[8] Edgar Seemann, Mario Fritz, and Bernt Schiele.
Towards robust pedestrian detection in crowded image sequences.
In IEEE Conference on Computer Vision and Pattern Recognition **CVPR'07**, Minneapolis, MN, June 2007.

[7] Mario Fritz, Geert-Jan M. Kruijff, and Bernt Schiele.
Cross-modal learning of visual categories using different levels of supervision.
In International Conference on Computer Vision Systems **ICVS'07**, Bielefeld, Germany, March 2007.

[6] Mario Fritz and Bernt Schiele.
Towards unsupervised discovery of visual categories.
In Pattern Recognition **DAGM'06**-Symposium, Berlin, Germany, September 2006.

[5] Mario Fritz, Bastian Leibe, Barbara Caputo, and Bernt Schiele.
Integrating representative and discriminant models for object category detection.
In IEEE International Conference on Computer Vision **ICCV'05**, pages 1363–1370, Beijing, China, October 2005.

[4] Mark Everingham, Andrew Zisserman, Christopher K. I. Williams, Luc J. Van Gool, Moray Allan, Christopher M. Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorko, Stefan Duffner, Jan Eichhorn, Jason D. R. Farquhar, Mario Fritz, Christophe Garcia, Tom Griffiths, Frederic Jurie, Daniel Keysers, Markus Koskela, Jorma Laaksonen, Diane Larlus, Bastian Leibe, Hongying Meng, Hermann Ney, Bernt Schiele, Cordelia Schmid, Edgar Seemann, John Shawe-Taylor, Amos J. Storkey, Sandor Szedmak, Bill Triggs, Ilkay Ulusoy, Ville Viitaniemi, and Jianguo Zhang.

The 2005 pascal visual object classes challenge.
In Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop **MLCW'05**, pages 117–176, Southampton, UK, April, 2005.

[3] Mario Fritz, Eric Hayman, Barbara Caputo, and Jan-Olof Eklundh.
The kth-tips database, 2004.
http://www.nada.kth.se/cvap/databases/kth-tips

[2] Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh.
On the significance of real-world conditions for material classification.
In European Conference on Computer Vision **ECCV'04**, pages 253–266, Prague, Czech Republic, May 2004.

[1] Matthias Zobel, Mario Fritz, and Ingo Scholz.
Object Tracking and Pose Estimation Using Light-Field Object Models .
In Vision, Modeling, and Visualization **VMV'02**, pages 371–378, Erlangen, Germany, November, 2002.